(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

CORRECTED VERSION

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 November 2001 (29.11.2001)

PCT

(10) International Publication Number
WO 01/090951 A2

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/US01/16375

(22) International Filing Date: 18 May 2001 (18.05.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/205,489 19 May 2000 (19.05.2000) US

(71) Applicant *(for all designated States except US)*: **THE BOARD OF TRUSTEE OF THE LELAND STANFORD JUNIOR UNIVERSITY** [US/US]; 900 Welch Road, Suite 350, Palo Alto, CA 94034-1850 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **HERZENBERG, Leonard, A.** [US/US]; 900 Welch Road, Suite 350, Palo Alto, CA 94034-1850 (US). **MOORE, Wayne** [US/US]; 900 Welch Road, Suite 350, Palo Alto, CA 94025-1850 (US). **PARKS, David** [US/US]; 900 Welch Road, Suite 350, Palo Alto, CA 94025-1850 (US). **HERZENBERG, Leonore** [US/US]; 900 Welch Road, Suite 350, Palo Alto, CA 94025-1850 (US). **OI, Vernon** [US/US]; 900 Welch Road, Suite 350, Palo Alto, CA 94025-1850 (US).

(74) Agent: **GOLD, Darren**; Howrey Simon Arnold & White LLP, 750 Baring Drive, Houston, TX 77057 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

(48) Date of publication of this corrected version:
18 September 2003

(15) Information about Correction:
see PCT Gazette No. 38/2003 of 18 September 2003, Section II

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/090951 A2

(54) Title: AN INTERNET-LINKED SYSTEM FOR DIRECTORY PROTOCOL BASED DATA STORAGE, RETRIEVAL AND ANALYSIS

(57) Abstract: The present invention is related to databases and the exchange of scientific information. Specifically the invention disclosed a unified scientific database that allows researchers to easily share their data with other researches. The present invention also allows for the ease of data collection, annotation, storage, management, retrieval and analysis of scientific data through and into the database. In addition, it allows for archival storage and retrieval of data collected directly from laboratory instruments to ensure data consistency for patent and other purposes. It also allows for ease of sharing data between laboratories in remote locations. The present invention also supports the automated creation of experimental protocols.

AN INTERNET-LINKED SYSTEM FOR DIRECTORY PROTOCOL BASED DATA STORAGE, RETRIEVAL AND ANALYSIS

SUMMARY

The present invention is related to databases and the exchange of scientific information. Specifically the invention disclosed a unified scientific database (IBRSS) that allows researchers to easily share their data with other researches. The present invention also allows for the ease of data collection, annotation, storage, management, retrieval and analysis of scientific data through and into the database. In addition, it allows for archival storage and retrieval of data collected directly from laboratory instruments to ensure data consistency for patent and other purposes. It also allow for ease of sharing data between laboratories in remote locations. The present invention also supports the automated creation of experimental protocols.

BACKGROUND

I. Fluorescent Activated Cell Sorting (FACS)

Flow cytometry is a technique for obtaining information about cells and cellular processes by allowing a thin stream of a single cell suspension to "flow" through one or more laser beams and measuring the resulting light scatter and emitted fluorescence. Since there are many useful ways of rendering cells fluorescent, it is a widely applicable technique and is very important in basic and clinical science, especially immunology. Its importance is increased by the fact that it is also possible to sort fluorescent labeled live cells for functional studies with an instrument called the Fluorescence Activated Cell Sorter (FACS).

Flow cytometry has always been computerized because without computers the data analysis would be infeasible. As flow cytometry has matured, the importance of combining flow data with data from other sources has become clear, as has the need for multi site collaborations, particularly for clinical research. This lead to our interest in developing methods for naming or identifying flow cytometry samples, reagents and

instruments (among other things) and in maintaining a shared repository of information about the samples etc.

Flow cytometry was revolutionized in the late 1970s with the introduction of monoclonal antibodies that could be coupled to a fluorochrome and used as FACS reagents. However, nomenclature for these reagents has been a hodgepodge, in spite of the fact that monoclonals are useful precisely because they can be uniquely and accurately named, i.e., the antibody produced by a clone is always the same whereas naturally produced sera are highly variable. Our work in capturing the experimental semantics of FACS experiments made it clear that we needed at least a local nomenclature and underscored the value of a global nomenclature for FACS data and monoclonal antibodies, which are useful in many fields beside flow cytometry.

II. DNA Arrays

During the past decade, the development of array-based hybridization technology has received great attention. This high throughput method, in which hundreds to thousands of polynucleotide probes immobilized on a solid surface are hybridized to target nucleic acids to gain sequence and function information, has brought economical incentives to many applications. See, *e.g.*, McKenzie, *et al., Eur. J. of Hum. Genet.* 6:417-429 (1998), Green *et al., Curr. Opin. in Chem. Biol.* 2:404-410 (1998), and Gerhold *et al., TIBS*, 24:168-173 (1999).

III. Gels

Gel electrophoresis is a standard technique used in biology. It is designed to allow sample to be pulled through a semisolid medium such as agar by an electro-magnetic force. This technique allows for separation of small and macromolecules by either their size or charge.

IV. Prior Art

Although there are wide variety of tools that purport to help scientists deal with the complex data collected in today's laboratories, virtually all of these so-called Laboratory Information Systems (LIMS) or Electronic Laboratory Notebook systems (ELNs) approach data collection and management from the perspective of final data

output and interpretation. None of these systems addresses the basic needs of the bench scientist, who lacks even minimal tools for automating the collection and storage of data annotated with sufficient information to enable its analysis and interpretation as a study proceeds.

The absence of automated support for this basic laboratory function, particularly when data is collected with today's complex data-intensive instrumentation, constitutes a significant block to creative and cost-effective research. Except in very rare instances, the study and experiment descriptions that scientists need to interpret the digitized data these instruments generate are stored in paper-bound notebooks or unstructured computer files whose connection to the data must be manually established and maintained. The volatility of these connections, aggravated by turnover in laboratory personnel, makes it necessary to complete the interpretation of digitized data as rapidly as possible and seriously shortens the useful lifetime of data that could otherwise be mined repeatedly.

In addition, because paper notebook or unstructured computer information is difficult to make available to other investigators, particularly at different sites or across time, laboratories that would like to make their primary data or their specific findings available to collaborators or other interested parties are unable to do so. Thus, although computer use now facilitates many aspects of research, and although the Internet now makes data sharing and cooperative research possible, researchers are prevented from taking full advantage of these tools by the lack of appropriately tailored computer support for integrating and accessing their work.

Finally, because the minimal computerized support for research that currently exists has developed piecemeal, usually in response to needs encountered during collection of particular kinds of data, no support currently exists for providing lateral support to integrate different types of data collected within an overall study. For example, although automated methods for collecting, maintaining and using DNA microarray data are now becoming quite sophisticated, the integration of these data with information about the source of the material analyzed, or with data or results from FACS or other types analyses done with the same material, is largely a manual task requiring recovery of data and information stored on paper or in diverse files at diverse locations that are often known only to one or a small number of researchers directly concerned with the details of the project. In fact, it is common for individual bench scientists to repeat experiments

sometimes several times because key information or data was "misplaced" or its location lost over time.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a diagram of the flow of information in a biological experiment

Fig. 2 is a diagram of a directory archival system

Fig. 3 is a diagram of information flow from instruments to and from the database (IBRSS) in one embodiment of the present invention

Fig. 4 is a diagram of information flow from instruments, analysis programs, remote databases, and other software and to the central database in one embodiment of the present invention.

Fig. 5 is a the hierarchical structure of a single study

DETAILED DESCRIPTION

The present invention will be best understood from the point of view of a laboratory worker using the invention. The invention may allow the user to simplify laboratory work by allowing interactive automation of much of the work with the use of a computer. The work that may be performed by the present invention may be able to make the researcher more efficient. The steps of the laboratory process the invention may address is collecting, sharing, retrieving, analyzing, and annotating data. Although the present invention has equal application to the storage of any data type, one embodiment relates to the storage of data associated with a biological sample data.

The first step the researcher may perform is to define a study **501**. A study may be defined as the overall goal of the research the researcher may wish to attain. In the normal course of science a researcher creates experiments to perform the research in the study. The study may contain protocols that capture the hypothesis to be tested and the factors that go into them, including subjects, treatments, experiments, samples and the study timeline. In addition, the study may contain data and information collected in

experiments that are part of the study. This may create a parent study node under which information and data pertaining to the study may be kept in child nodes.

The present invention may allow a researcher to create experiments and experimental protocols **502** and **503** that may become part of the overall study. The experiment may contain protocols that acquire information to define the subset of subjects for which the data may be collected, the set of samples to be obtained from the subjects, and the analytic procedures and data collection instruments used to analyze the samples. The experiment protocol may become a child node of its parent study.

As a typical researcher does today, the researcher using the present invention also may obtain data **504** and **505** for each study and experiment he performs. The data may be collected each time the researcher performs the same experiment protocol. The data may also contain protocols designed to acquire annotation information to define the subdivision (aliquotting) and the treatment (reagents and conditions) for a set of samples for which data may be collected by a single analytical method (usually a single instrument). Researchers then analyze data they obtain, and the researcher using the present invention may analyze the collected data. This analysis may stored as a child-node of the data or the annotation of the data **506** and **507**.

When the analysis is complete, the present invention may create Internet addresses for all of the results of the individual analyses and for the data sets created. These may be child nodes **508** and **509** of the data or experiment information. Thus, the present invention allows the user to possess unique web addresses for any of the data or analysis results that he may wish to include in a publication. The study, experimental protocol, data collection, and analysis results, may be stored as described in FIG. **5.**

The study and the experiment are still the touchstone of research science. The present invention may allow the researcher to interactively create protocols for studies and experiments. The protocol creators may use wizards to ease the researcher's creation of the protocols. The researcher may invoke a protocol creator/editor on a computer. The computer may provide the researcher with a list of possible studies or experiments the researcher may wish to perform. The computer may also provide the ability for the researcher to create an entirely new type of study or experiment. After the type of study

or experiment is chosen, the researcher may then be a given the option of how to set up the experiment.

Several types of possible studies, experiments and options are listed here, however the person ordinarily skilled in the art will realize how to put other types of experiments into the present invention. The types of experiments that will be described in this application specifically are clinical and basic studies and FACS and electrophoresis gel experiments. Other types of data that can be similarly stored and used within the database include DNA microarray data and clinical data. The clinical data may include red blood cell counts and RBC, MCV, MHC, MCHC, and potassium levels or may include observational data such as blood pressure, temperature, types of drugs taken, race, age, etc.

An example of a study may be a clinical study. The study may be designed to test one or more hypotheses. An example of a hypothesis may be testing whether the number of CD8 T cells is correlated with the erythrocyte volume.

In the study, HIV-infected patients may be recruited on the basis of meeting a series of entry criteria. Examples of such criteria are:

1) information collected directly by interviewing the patient

2) results of clinical analyses such as erythrocyte counts

3) results of FACS analyses such as number of CD4 T cells

Experiments in the study may be conducted on samples from patients to determine whether the patient meets the entry criteria for the study. In this case, information and experiment results for each potential study entrant may be stored in the study. The study may contain experiments such as staining cells from the patients with antibodies that reveal cells that express surface CD4 and analyses such as those that enumerate the number of cells expressing CD4. Relevant information about the subjects (patients) in the study may be passed from the study to protocol wizards that may help the user define the contents of experiments such as which samples from which subjects may be examined. The study may also allow the user to select from model protocols for the experiment to define types and the amounts of the FACS reagents that may be used. For example, once information for a subject is entered into the study, the study subject may appear on a list from which the user chooses the samples to be examined in an experiment.

The study may also specify that the protocol automatically send data that is collected to analysis programs and provide necessary information to enable the automated analysis and to return specified results of the analysis to the study. Similarly, when these data are returned, the study may be triggered to specify automated analyses that return further digested results to the study. One result of this process may be the automatic identification of subjects that qualify for further study by determining that the study criteria are met, such as the subjects' erythrocyte counts and CD4 counts are within the specified ranges. Further, the automated analysis may include the returning of FACS plots comparing CD4 and CD8 levels, the returning of charts with each subject's mean levels of CD4, CD8, erythrocyte counts, or other specified variables. The automated analyses may also specify the performance of statistical procedures and the return of results of these analyses. In addition, the study may have methods for summarizing and displaying results of analyses. Finally, the study may track samples to determine whether required experiments were performed and specified data returned and may contain information about the physical location of stored samples, the amount of the sample that has been used, the treatment of the sample.

A basic research study may contain samples from mice, information about the genetic makeup of the mice and references to genome other databases relevant to the mice. It may also contain information about the treatments that individual or groups of mice were given or may be given during the experiment and about the drugs or other materials with which the mice were or may be treated. The study may also contain the timeline for treatment and, as above, define protocols and automated analyses for collected data.

A FACS experiment in a study comprises staining cells with various fluorescent antibodies and running and possibly collecting cells through a cell sorter. The wizard may help the experimenter create his experiment by creating a suggested protocol for him to follow. The wizard or other interactive device may ask the researcher how many different stains he wishes to use to mark various structures. These stains may, but do not necessarily need to be stains for different structures. Typically the stains may be fluorescent conjugated antibodies. The user may then inform the protocol creator which structures he wishes the stains to mark and the wizard may respond with an offer of a series of "option" lists from which the user may select the type of cells and the specific

reagents to be used in the experiment. Option lists may be generic types of cells or cells and samples specified in the parent study to which the experiment belongs.

The wizard then may ask the researcher which FACS machine he plans to use. Each FACS machine may be equipped with different lasers or light filters enabling different FACS machines to collect data for antibodies labeled with different fluorescence "colors". The wizard may then determine whether the FACS machine specified by the user is able to take data for the fluorescent reagents selected in the protocol. Alternatively, the wizard my suggest which of the FACS machines available to the user can be used. In either case, the wizard may then assist the user in scheduling an appropriate analysis time period on an appropriate FACS machine.

Finally, the protocol creator may use combinatorics or other procedures to define the reagent and cell sample combinations that the user may have to pipet (add to tubes) to complete the experiment and create a protocol for the researcher to follow. This protocol may specify the control tubes that are required and provide the concentrations and amounts of antibodies to use, the dilutions of the antibodies, the various steps to perform, the various centrifugations to perform, and the FACS to operate. Typically a control tube may be suggested for each antibody employed in the study. Further a blank control tube for each separate organism may be suggested to determine autofluorescence.

The reagents used by the protocol may have attributes associated with them. These attributes may include the reagent's distinguished name, Clone ID, Common name, Specificity, Titre, Fluorochrome Name, Fluorochrome Lot number, and concentration. The user may be prompted to select the reagents used through a "Reagent Palette". Such a palette may contain a catalog of reagents in stock, pre-determined sets of reagents typically used in similar protocols, and an ability for the user to enter a new choice of reagents for the experiment.

The protocol creator may also perform various tasks behind the scenes to create a valid protocol for the researcher, to call for pre-packaged analyses, to check data quality during data collection, and to display the information about the reagents and cells in a sample at the time of data collected or any other time.

The protocol editor may be tied to a database to enhance its, as well as the researcher's efficiency. In the previous example, several items may be used from the database to create the FACS protocol. For example,

1) The database may hold data for the fluorescent recognition abilities of all of the FACS machines available to the user. This may allow the protocol editor to select only those reagents that are available to the user and can be viewed by the FACS chosen by the user. There are a wide variety of possible combinations of possible reagent choices that can be selected. Specifically, there may be n!/(n-k)!k! possible reagent choices where n is the total number of fluorescent "colors" that for which the FACS can collect data and k is the number of stains used in the FACS experiment. However, this number is restricted because not all reagents are available in all colors.

The present invention may provide a novel way to enhance the effectiveness and speed of the selection of the reagent combination by applying well know combinatorial techniques and depth-first search in a new way to this biological problem. This may be performed by selecting one reagent at a time recursively. If the most recently added reagent cannot be used with the current set, then that reagent may be removed from the list of suggested reagents. The algorithm may run until a set of usable reagents is determined.

2) The protocol creator may also consult laboratory databases to determine how much of each reagent may be available to the user. If the protocol creator finds that the amount of reagent available is below a pre-set threshold, it may automatically indicate the reagent shortage and suggest another combination to be used. The protocol creator may also consult the database as to the effectiveness of each stain to bind to the type of cell being used. It may then use a greedy or any other algorithm (such a s the ones suggested to select reagents combinations) to select an optimal set of stains to be used in the experiment. Other factors may also be taken into this optimization including the price of the reagents, the temperature compatibility of the reagents in a given combination, and the resolution possible for target cell surface or internal markers when stained with the selected reagent

combination. This may be performed using a scoring function that provides a score for each of the factors in selecting the reagents.

3) The protocol creator may suggest the layout of the wells, tubes, or containers used to perform the experimental protocol. The layout may depend on the proximity of like samples, like reagents, and controls. The layout may also be created to minimize the movement of the person undertaking the protocol. Such an instance would be when several tubes require the same regent cocktail. In this case, it would be of benefit to have those wells, tubes, or containers located near one another. The protocol editor may also suggest the creation of reagent cocktails when several reagents with the same proportions are needed in various wells, tubes, and containers. The reagent cocktails may be designed by determination of like reagents used in multiple wells. This determination may be through linear programming or another optimization routine designed to minimize the number of pipeting steps or any other experimental concern such as time, cost, or ease. The constraints for such a linear programming model may include any of the aforementioned factors contributing to experimental time, ease, or cost.

4) The protocol creator may also suggest the use of different FACS machines that are capable of performing the experiment because either the FACS machine may be cheaper to operate or the cost of the reagents for that FACS machine may be cheaper. The protocol creator may also anticipate what type of data may be collected and may prepare table and charts to be filled in after the experimental data is collected. One method of creating charts may be to create 2-axes graphs for all the pairs of data that the protocol is expected to collect.

After a protocol is created and/or used, the protocol creator may then allow the user to store and re-use the protocol in the database under the current study or any other study the scientist wishes to use the protocol for. Once data collection for a sample is complete, the protocol creator may cooperate with the data collector to couple the collected data with the annotation information (reagents, cells, treatments) known to the creator and may send the coupled data and annotations to the database for permanent storage and archiving. Once the data collection for a full experiment is complete, experiment-related information

(standards, machine conditions, etc) may be sent to the database to be coupled with the sample data and annotation. These couplings may be accomplished by storing the data separately from the annotation data and associating these items permanently by use of non-volatile pointers or some other means. The parent study may also be informed of the completion of the experiment and the location of the output from the experiment (protocol and data collection).

After the scientist creates the protocol, he is now able to perform the protocol and conduct the experiment. This experiment may create data that may automatically be captured by the database, coupled with the annotation information in the protocol, transferred from the machine used to collect the data (FACS, in the example above) directly to the proper location for the particular experimental data. This can be performed in several ways, including the use of LDAP, XML and XSL style sheets. Analysis programs may automatically perform preliminary analysis specified by the protocol or elsewhere. The protocol editor may determine the nature of data and may inform the analysis program the type of data that is represented. The data types may include nominal, ordinal, or continuous that are either dependant or independent variables. The variables may also be crossed or nested. These analyses may be informed by the annotation and possibly other information associated with the data (such as data type) collected for each sample. Results from these preliminary analysis may be stored and associated with the collected data and be locatable via an experiment data tree that may be available for the experimenter to view. For FACS analysis the collected and annotated data may automatically be sent to a FACS data analysis program such as FloJo or CellQuest. Once FACS analysis begins, the analysis software may suggest possible gating strategies with the use of clustering algorithms or other artificial intelligence techniques. Further gating data may be displayed using the annotations from the protocol editor to determine the labeling of the axes of the displayed data. The data also may be sent for analysis to a statistics analysis package such as JMP (from the SAS Institute). The data may be automatically processed to determine such statistics as median attribute values and standard deviations of attribute values.

As with any other scientific or engineering method, Gel electrophoresis may also be incorporated into the current system of protocol development. For instance, the protocol creation wizard may prompt the user to select/input the type of gel that is to be

run. These gels may include a Northern or Southern blot. Further, the wizard may prompt the user to input the number of lanes in the gel and select the sample is to be placed in each lane. The sample may be defined at the protocol level or may be selected from in a list generated from information already entered into the study to which the experiment protocol belongs. Further, the protocol creation wizard, possibly informed by the study, may prompt the user to determine which type or types of standard controls, such as ladders, are going to be used in the experiment. The protocol wizard may suggest the lanes that each specimen should be placed in according to rules pre-defined for the type of gel and sample in the experiment.

After the experiment is completed, the user may bring the gel to an instrument for automated or manual data collection. For instance, the user may bring the gel to an ultra-violet gel reader connected to a computer. The reader may take a picture of the gel and send a digitized version, coupled with the protocol information that describes the sample and the experiment, to a central data store for archiving. The gel reader may then send the digitized picture to an analysis program. Alternatively, the data in the data store may be sent at the user's request, to the analysis program. This analysis program may determine the size of each fragment found in the gel by comparing their positions to the positions of the ladder. The results of the analysis may then archived in the database for later retrieval, further analysis or abstraction into summaries in the parent study. The parent study may also be informed of the completion of the experiment and the location of the output from the experiment (protocol and data collection).

There are several experimental models which may be incorporated into the database. These models may be selected by the user to provide the protocol creator what type of experiment to create. The experimental models may include:

1) Crossing Model: Many experiments are essentially combinatorial, i.e., this set of reagents or reagent cocktails is applied to each sample in a group of samples. Typically it may correspond to some N X M grid of wells in the staining plate. An experiment might have 1 or more of these repeated sets of reagents.

2) Titration Model: The user may specify a target sample and a reagent and then a range of dilutions 2, 4, 8... or 10, 20, 50, 100 being typical. The layout

of the dilution may be as a single column, a single row, or otherwise on the plate or other type of container.

3) Screening Model: The user may specify a reagent cocktail and a large number of samples which are quasi-automatically named.

4) Fluorescence Compensation Controls Model: For each dye (or dye lot) which occurs in an experiment model, the user or protocol editor may specify a sample to be used as a control. Usually the control will be one of the samples which is stained with the reagent.

5) Unstained Controls Model: The user or protocol editor may define an unstained or negative control for a protocol involving staining. Unstained controls and fluorescence compensation controls may be coupled in a together in a single experimental protocol to create a population of suitable controls.

The protocol editor may create a GUI representing the wells, tubes, or other containers holding the reagents and samples. The user may be able to "drag and drop" the sample or reagent to another well, tube, or container to alter the experimental protocol the user created or the protocol creator suggested.

After the study is completed the software may test the hypothesis stated in the study protocols. The hypothesis may be test by combining the statistical information gathered during the experimental protocols and determining if they fit the hypothesis. This determination may be done manually by viewing the data or automatically by allowing the data to be analyzed by a data analysis package such as JMP. In one embodiment, JMP may automatically analyze the data that may be specified by the user when the user creates an experimental protocol with the appropriate wizard. The wizard may then associate the expected data with the study node with so that the hypothesis may automatically be tested.

The database may allow access to the data for several purposes. First, the user may be able to provide hyperlinks to collected data and experimental protocols so that others may access the data and protocols. Others that would access the data may include collaborators, reviewers, and others reading published articles containing hyperlinks to the data. Second, the database may act as a cell surface expression library enabling people such as researchers and clinicians to facilitate diagnosis and definitions of new

conditions by comparing the data from the database with locally collected data. Other uses of this database would be obvious to those skilled in the art.

The database may be constructed using any known database technique including the use of LDAP directories and protocols, XSLT style sheets, and XML documents. The database may be at a centralized site remote to the experimenter. The experimenter may send or receive information between his computer and the database via the Internet or any other communication means. LDAP is a "lightweight" (smaller amount of code) version of DAP (Directory Access Protocol), which is part of X.500, a standard for directory services in a network. The present invention may put these to unique uses in the scientific arena. In essence, the style-sheet transformation language (XSLT) defines the transformation of the original input (XML) document to "formatting objects" such as those included in HTML documents. In a traditional style sheet, these are then rendered for viewing. However, the XSLT transformation grammar can also be used to transform XML documents from one form to another, as in the following examples:

a) **Loading directories.** XSLT may be used to transform an XML file generated by any data processing application to an XML representation of a directory (sub)tree, i.e., to extracting directories entries from the XML document. The ability to use XSLT for this transformation greatly simplifies the creation and maintenance of LDAP or other directories that serve diverse information derived from distinct sources (e.g, FACS instruments and genome data banks) that generate different types of XML documents. In essence, using XSLT removes the necessity for writing distinct Java code to construct the directory entries for each type of document. Instead, appropriate "directory styles" can be defined for each document type and a single Java program can be written to process all XSL-transformed documents into the directory tree.

b) **Re-indexing directory entries.** Existing documents may be readily re-indexed based on any desired elements or attributes present in the XML documents simply by changing the XSLT style sheet. Changes in the directory schema may be required for extensive indexing changes but could also be driven by an XML representation of the appropriate schema.

c) **Cataloging new documents.** A new type of document can be cataloged simply by creating an appropriate XSLT style sheet and modifying the directory schema if necessary, as above.

d) **Cataloging from arbitrary XML documents.** A default XSLT directory style sheet can be created to extract a pre-defined set of indexing elements included in arbitrary XML documents. This would enable creation of the corresponding directory entries for these indexing elements.

e) **Passing information from XML files to analytic or other programs:** XSLT can be used to transform a subset of the information in an XML file so that it can be read by a program that takes XML input in a particular format. In addition, XSLT can launch the program and pass the result of the transformation during the launch. For example, using XSLT stylesheets, we can launch an analysis application by transforming an XML file containing the results of a directory search to an application-readable file containing URLs for the data and appropriate annotation information for the analysis. This option can be made available for all co-operating applications and need not be restricted to FACS data.

f) **Creating data displays.** XSLT style sheets can be used to change the form of a document. For example, they can be used to extract the results of analyses and display them as values in the rows or columns of a table.

As indicated above, XSLT and other capabilities may be used to store analysis output along with the primary data and annotation information. Alternatively, other developed fully cooperating applications may be used to analyze of FACS and other data.

A major advantage of LDAP is the availability of LDAP servers and client toolkits. Standalone servers and LDAP to X.500 gateways are available from several sources. LDAP client libraries are available for the C language from Univ. Michigan and Netscape and for the Java language from Sun and Netscape.

Secondly, LDAP is a standard that is directly utilized by the clients and makes it possible for all clients to talk to all servers. In contrast, SQL standardization may be more apt with transportability of programmers and database schema than interoperability of databases.

The X.500 information model is extremely flexible and its search filters provide a powerful mechanism for selecting entries, at least as powerful as SQL and probably more powerful than typical OODB. The standard defines an extensibleObject that can have any attribute. Furthermore, some stand-alone LDAP implementations permit relaxed schema checking, which in effect makes any object extensible. Since an attribute value may be a distinguished name, directory entries can make arbitrary references to one another, i.e., across branches of the directory hierarchy or between directories.

Finally, some LDAP and X.500 servers permit fine grained access control. That is to say, access controls can be placed on individual entries, whole sub trees (including the directory itself) and even individual attributes if necessary. This level of control is not available in most existing databases.

One example of an LDAP directory is organized in a simple "tree" hierarchy consisting of the following levels:

1) The "root" directory (the starting place or the source of the tree), which branches out to

2) Countries, each of which branches out to

3) Organizations, which branch out to

4) Organizational units (divisions, departments, and so forth), which branches out to (includes an entry for)

5) Individuals (which includes people, files, and shared resources such as printers)

This example tree structure of an LDAP directory is illustrated in Figure 2. The parent node of the tree is the root node 201. The children of the root directory are country nodes 202.1 and 202.2. Each country node can have child organization nodes such as organization nodes 203.1 and 203.2 (children of country node 202.2).

Below the organization level are organization group nodes such as nodes and 204.3 which are children of organization node 203.2 Each group can have children nodes representing individuals such as group node 204.3 having children nodes 205.1,205.2, and 205.3.

In a network, a directory tells you where in the network something is located. On TCP/IP networks (including the Internet), the Domain Name System (DNS) is the directory system used to relate the domain name to a specific network address (a unique location on the network). However, sometimes the domain name is not known. There, LDAP makes it possible to search for an individual without knowing the domain.

An LDAP directory can be distributed among many servers. Each server can have a replicated version of the total directory that is synchronized periodically. An LDAP server is called a Directory System Agent (DSA). An LDAP server that receives a request from a user takes responsibility for the request, passing it to other DSAs as necessary, but ensuring a single coordinated response for the user.

The present invention contemplates extensions and modifications to LDAP protocols to make them usable not just as directories, but to also provide data itself. The present invention takes advantage of hierarchical levels of LDAP already established by the International Standards Organization (ISO) and uses those organizations to provide a first level of uniqueness to the biological sample to be named.

Referrals mean that one server which cannot resolve a request may refer the user to another server or servers which may be able to do so. During a search operation any referrals encountered are returned with the entries located and the user (or client) has the option of continuing the search on the servers indicated. This allows federation of directories which means that multiple LDAP/X.500 servers can present to the user a unified namespace and search results even though they are at widely separated locations and the implementations may actually be very different.

The Java Naming and Directory Interface (JNDI) is a standard extension to the Java language introduced Java Naming and Directory Interface by Sun. It includes an abstract implementation of name construction and parsing that encompasses the X.500 name space (among others), and an abstract directory that is essentially the X.500 information and functional models. Specific implementations (service providers[13]) are available for LDAP, Network Information Server (NIS) and even the computers own file system.

JNDI may remove many of the limitations of LDAP as an OODB by providing a standard way to identify the Java class corresponding to a directory entity and instantiate it at runtime. It also allows storage of serialized Java objects as attribute values. Sun has proposed a set of standard attributes and objectClasses to do this.

When represented as a string (essentially always with LDAP) an X.500 distinguished name is a comma separated list of attribute value pairs and is read from right to left. If a value contains special characters such as commas it must be quoted and in any case initial and final white space around attributes or values is ignored. For example, "cn=Wayne Moore, ou= Genetics Department, o=Stanford University".

Location names may have as their root (right most) component the countryName or c attribute with the value being one of the ISO standard two letter country codes, for example c=US. Such names can be further restricted by specifying a stateOrProvinceName abbreviated st and a locality abbreviated l, for example "l=San Francisco, st=California, c=US".

Organizational names may have as their root the name (registered with ISO) of a recognized organization and may be further qualified with one or more organizational units, for example "ou=Department of Genetics, ou=School of Medicine, o=Stanford University".

Domain names as used by the Domain Name Service (DNS) are represented with the dc attribute, for example, "dc=Darwin, dc=Stanford, dc=EDU".

Names of persons. There are two conventions for naming people. The older uses the commonName or cn attribute of the Person objectClass but these are not necessarily unique. Some directories use the userId or UID attribute of inetOrgPerson, which is unique. Since uniqueness is important for scientific applications the latter may be used. The remainder of a person's dn is usually either an organizational or geographic name, for example "uid=wmoore, o=Stanford University" or "cn=Wayne Moore, l=San Francisco, st=California, c=US".

Examples of encapsulating and extending existing nomenclatures:

1. **Gene loci,** for example, "locus=Igh-1, o=Professional Society or locus=New, cn=Leonard Herzenberg, ou=Department of Genetics, ou=School of Medicine, o=Stanford University".

2. **Gene alleles,** for example, "allele=a, locus=Igh-1, o=Professional Society or allele=1, locus=127, ou=Department of Genetics, o=Stanford University".

3. **CD antigens,** for example, "specificity=CD23, o=Human Leukocyte Differentiation Workshop".

4. **Literature references** in the scientific literature are essentially achieved the benefits of distinguished names without an explicit central authority. However representing them as distinguished names may facilitate mechanical processing. For example, "title="A Directory of Biological Materials", volume=1999, o="Pacific Symposium on Biocomputing". A true directory of such literature references would be of obvious value over and above the current unique naming systems in some of the current literature archives.

5. **New nomenclature schema.** The following schemas arose from work on storing information about flow cytometry data in directories.

6. Monoclonal antibodies are distinguished by cloneName or clone which is unique within the parent entity which must be an investigator or organization.

7. Lymphocyte differentiation antigens, a thesaurus of the target specificities of monoclonal antibodies. Would include but not be limited to the official CD names.

8. FACS instruments are distinguished by the cytometer attribute which must be unique with respect to the organization parent, for example, "cytometer=Flasher II, ou=Shared FACS Facility, o=Stanford University".

9. FACS experiments are distinguished by the protocolIdentifier or protocol attribute which must be unique with respect to the parent which may be a person, and instrument or and organization or some combination, e.g., "protocol=1234, cytometer=Flasher, uid=Moore, ou=Shared FACS Facility, o=Stanford University".

10. FACS samples are distinguished by a unique protocolCoordinate which must be unique within the parent FACS experiment, e.g., "coord=A12a, protocol=12345, cytometer=Mollusk, ou=Shared FACS Facility, o=Stanford University".

Therefore, using LDAP any object, such as a monoclonal antibody, may be named relative to the unique distinguished name of an investigator or organization. That means that unique identifiers can be assigned to biological materials early in the scientific process and thus facilitate professional communication both informal and published. In the future, investigators who have this distinguished name can identify the material unambiguously via the unique name. If a directory services is maintained, an investigator can determine if the sample has been given an official name, if it has been shown to be equivalent to another entity or if it has been cited in the literature.

Directory searches may also be a tool available in the database. Information may be promoted upward from the documents into the directory for searching and no searching is done within the documents. However, since XQL or Xpath allows searches to proceed downwards from the directory, a search application may use the LDAP search functions to retrieve a set of candidate XML documents (based on their directory attributes) and then may use XQL or Xpath to further refine this set. To facilitate XQL or Xpath use, a unified interface may be provided that would largely make the differences in search strategies transparent to the user. The user then may be able to select (search and retrieve) for items within the document that are not reflected in the directory or may extract elements from these documents, e.g., samples from a set of experiments.

The instruments may be responsible to collect, annotate and export the collected experimental data. The instruments may annotating it with information generated during the data collection, and for transmit the annotated primary data to the LDAP server for storage in the database in association with the appropriate XML-encoded experiment and study descriptions. The following modules may be used to perform these functions:

a) **Set-up module(s)** – automate aspects of instrument set-up and standardization; record and visualize relevant instrument information; acquire and respond to user input

b) **Data collection module(s)** – collect primary (instrument-generated) data for the aliquots of each sample; visualize protocol information to facilitate data collection; acquire and respond to user input; record machine condition and user comments specific to each data collection.

i) adapt and interface the data collection modules to specific machines (e.g., various FACS, imaging and DNA-array data readers) to provide full functionality for data collection.

ii) For instruments that do not provide/permit direct access to machine control and data collection, use additional modules that may enable manual entry of machine information and "point-and-click" association of primary data collected for each sample aliquot with the protocol information for that aliquot.

c) **Extension of the FACS document type** - include new functionality such as instrument setup, auto-calibrator and quality control elements, tabulated transfer functions and operator commentary in the definitions of the FACS document type. Provisions for digests of the data files that are referenced and for digital signatures may also be made.

d) **Data transmission module(s)** – link (annotate) the primary data with protocol instrument-derived information; communicate authenticated (digitally-signed) primary data and its annotation linkages to the information store.

The central database may be a large scale (terabyte level), web accessible, central storage system coupled with small-scale volatile storage deployed locally in a manner transparent to the user. This system may store data and annotation information transmitted from the data collection system. In addition, it may catalog the stored data according to selected elements of the structured annotation information and may retain all catalog and annotation information in a searchable format. Wherever possible, industry standard formats for storing data and annotation information will be implemented. If no standard is available, interim formats may be used and may allow for translators to industry standards once the industry standards become available.

The database may capitalize on the built-in replication and referral mechanisms that allow search and retrieval from federated LDAP networks in which information can be automatically replicated, distributed, updated and maintained at strategic locations throughout the Internet. Similarly, because pointers to raw data in LDAP are URLs to data store(s), the database may capitalize on the flexibility of this pointer system to enable both local and central data storage.

The database may enable highly flexible, owner-specified "fine-grained" access controls that prevent unauthorized access to sensitive information, facilitate sharing of data among research groups without permitting access to sensitive information, and permit easy global access to non-sensitive data and analysis results.

a) Built-in access controls that may prevent release of unauthorized information from the system

b) Multi-level access controls that may allow data owners to specify which users, or classes of users, are permitted to retrieve individual data sets and/or to access individual elements of the annotation information during searches

c) User identity verification system that may be referenced by the access control system

d) Anonymous access to data and annotation information that owners may make available for this purpose

e) Security and encryption may be implemented to protect the information in the database itself as well in the communications between the central data repository and the remote locations.

The central database may also allow for the retrieval of annotated data sets (subject to owner-defined accessibility) via catalog browsing and/or structured searches of the catalog; The central database may also automatically verify authenticity of the data based on the data's digital signature. This function may be accomplished by launching internal and co-operating data analysis and visualization programs and transfer the data and annotation information to the program. Further the database may put the data and annotation information into published-format files that can be imported into data analysis and visualization programs that do not provide launchable interfaces.

The central database may also allow for retrieval of analysis output. This function may be accomplished by recovering/importing the link analysis output with primary and annotation data to provide access to findings via subject and treatment information that was entered at the study and experiment levels. This may allow the database to store and catalog output from co-operating analysis programs (within the limitations imposed by the capabilities of analysis programs that were not designed for this purpose). It may also

allow the database to use internal analytic modules and programs that may enable users to fully capitalize on the annotation information entered into the system.

A DIRECTORY OF BIOLOGICAL MATERIALS

WAYNE A. MOORE

Genetics Department, Beckman Center B007, Stanlord University, Stanford, CA 94305-53 18, USA

Systematic nomenclature has been an essential tool in biology since its emergence as a modem science. However, the method by which formal or official names are adopted, namely meetings by professional or governmental bodies, has not changed since Linnaeus.. The last decade has seen rapid advances in the standardization (X. 500, LDAP)and implementation of computerized directory services, including a global system of distinguished names. This paper is *a* proposal that the biomedical community adopt X. 500 as a standard for the machine representation of biological names. Adherence to such a standard would permit the sharing of essential information about research materials through directories. Adoption of unique names for biological materials facilitates collaboration by enabling investigators to exchange (vie email or electronic publication)unique identifiers for materials. An actively maintained directory of such materials would provide collaborators and future investigators with access to the primary data referenced by the literature, information about changes in nomenclature (for example adoption of a standard name by a professional society) and references, citations or hyperlinks to later work on the material. We are implementing such a directory of flow cytometry samples and the monoclonal antibody reagents used to prepare them. A minimal set of names and objects drawn from this effort is provided here as a concrete example.

1 Introduction

Flow cytometry[1] is a technique for obtaining information about cells and cellular processes by allowing a thin stream of a single cell suspension to "flow" through one or more laser beams and measuring the resulting light scatter and emitted fluorescence. Since there are many useful ways of rendering cells fluorescent, it is a widely applicable technique and is very important in basic and clinical science, especially immunology. Its importance is

increased by the fact that it is also possible to sort fluorescent labeled live cells for functional studies with an instrnrnent called the Fluorescence Activated Cell Sorter (FACS). At our FACS facility alone, we have processed millions of samples in the last 15 years.

Flow cytometry has always been computerized because without computers the data analysis would be infeasible. As flow cytometry has matured, the importance of combining flow data with data from other sources has become clear, as has the need for multi site collaborations, particularly for clinical research. This lead to OUT interest in developing methods for naming or identifying flow cytometry samples, reagents and instruments (among other things)and in maintaining a shared repository of information about the samples etc.

Flow cytometry was revolutionized in the late 1970s with the introduction of monoclonal antibodies[2] that could be coupled to a fluorochrome and used as FACS reagents. However, nomenclature for these reagents has been a hodgepodge, in spite of the fact that monoclonals are useful precisely because they can be uniquely and accurately named, i. e., the antibody produced by a clone is always the same whereas naturally produced sera are highly variable. Our work in capturing the experimental semantics of FACS experiments made it clear that we needed at least a local nomenclature and underscored the value of a global nomenclature for FACS data and monoclonal antibodies, which are useful in many fields beside flow cytometry.

There are many existing nomenclatures in biology and medicine that provide uniqueness by specifying a central registry, usually mediated by a professional society. Instead, to ensure uniqueness without global meetings, International Standards Organization (ISO) X.500 directory servers[3] achieve uniqueness with distinguished names (dn) that are assigned hierarchically. IS0 defines country names and registers organization names, e.g., "c=US" and "o=Stanford University" respectively. Governmental or non-governmental organizations then define how relative distinguished names are handed out, e.g., by state "st=Califomia, c=Us" or by organizational unit "ou=Genetics Department, o=Stanford University".

It is easy to represent traditional standard names within the X. 500 standard distinguished names: simply make them relative to the organization which defines them. Objects such as monoclonal antibodies can be named relative to the unique distinguished name of an investigator or organization. That means that unique identifiers can be assigned to biological materials early in the scientific process and thus facilitate professional

communication both informal and published. Later, investigators who have this distinguished name can identify the material unambiguously and if a directory services is maintained, determine if it has been given an official name, if it has been shown to be equivalent to another entity or if it has been cited in the literature. Thus I propose here, both for flow cytometry and as a general practice in biocomputing, the use of X. 500 nomenclature. At the Stanford Shared FACS Facility we are constructing a testbed for these concepts applied to flow cytometry, based on commercial LDAP directory servers.

2 Background

2.1 *Directories: X. 5002, LDAP v2 and v3*

X. 500[3] is the core of a set of standards adopted by the International Standards Organization (ISO) beginning in 1988, which defines what may be simply called directory service. A directory is fundamentally a database. Directories were originally defined in order to allow users and their agents to find information about people, typically their telephone number but possibly including postal address, email address and other information. This was extended to include documents, groups of users and network accessible resources such as printers and more recently databases. Three parts of the standard are of particular interest, the information model, the functional model and the namespace.

The X. 500 information model is very powerful and flexible. The standard defines entries which have a set of named attributes that can have one or more values and may be absent. Each attribute has a name and a type and each type has a name and a syntax which is expressed in Abstract Syntax Notation One (ASN. 1). By default the types case exact string, case ignore string, telephone number, integer, distinguished name and binary are recognized. Every entry must have an attribute object Class which defines what attributes are possible and which are required and may have an attribute aci (for access control information) which the server uses to control access to the entry. Object classes are hierarchical, i. e., a class can inherit attributes from a parent class and by defining new attributes extend it's scope

The entries in a directory are organized hierarchically. That is to say that any entry may have one or more subentries so that the whole structure may be visualized as a tree. At every node each subentry is identified by a value of one of its attributes called a relative

distinguished name (rdn) which must be unique within its level, for example "uid=wmoore". A distinguished name of a subentry is defined by concatenating its rdn with the dn of its parent entry which is likely to be itself a compound name, for example "uid=WmOore, ou=Shared FACS Facility, o=Stanford University". These distinguished names are the namespace mandated by X. 500.

The functional model defines a set of operations which may be applied to a directory: read, list, search, add, modify, delete (which are pretty much self explanatory) and bind, unbind and abandon which are used to establish the users credentials, end a connection to the server and cancel a running query respectively.

The search function starts from a root dn and finds all entities further down in the hierarchy which pass a search filter constructed from the "usual suspects", i. e., equal, less than, contains, sounds like etc. applied to the attributes of the entity. A search filter may of course test the objectClass attribute and return only entries of a particular type. Clients can specify searches which return all the attributes of each entry or only a selected set of attributes.

The protocol defined in X.500 for accessing the Directory Service Agent (DSA) is called Directory Access Protocol (DAP) and it runs on the Open System Interconnect (OSI) protocol stack which is also in its own right an IS0 standard. This fact as well as the complexity of the security mechanisms and abstract attribute encoding of the full protocol made it difficult to implement DAP on lightweight clients, i. e., PCs and Mats.

The complexity of an X. 500 directory client led to a desire for X. 500 lite or a Lightweight Directory Access Protoco[4,5] (LDAP)which would run on the TCP/IP protocol stack that is widely available on lightweight clients. LDAP adopts the X. 500 data model essentially intact. It simplifies the functional model by collapsing the read, list and search functions into a single search function with object, one level or sub tree scope respectively. It handles distinguished names as strings rather than the structured objects that DAP uses which transfers the responsibility for parsing them to the server. Conversely most of the responsibility for interpreting the attribute values reverts to the client. This results in some loss of robustness (because of weaker type checking) but relieves the client of the need to parse abstractly (ASN. 1) defined objects. LDAP returns the results as individual packets which allows lightweight clients to process result sets which they cannot store in memory. LDAP does not include much of the elaborate security and authentication mechanisms used

by DAP and also simplifies the search constraints to the maximum number of entries to return and maximum time to spend searching.

Unfortunately one X. 500 function known as referral was not included in LDAP v2. This allows one DSA to return to the client a referral which directs the client to try again on a different DSA. An LDAP v2 server is supposed to follow all referrals on behalf of the client and not return them to the client at all.

LDAP v2[5]was proposed to the Internet Engineering Task Force (IETF) as a draft standard but was not adopted due to its technical limitations. This lead to the effort to define a more acceptable version. Also in this period the utility of stand alone LDAP servers, i. e., servers which implemented the information and functional models directly rather than relying on a higher tier of X, 500 servers became clear.

LDAP v3[6] addresses the problems discussed above and was adopted by IETF in 1998 as a proposed standard for read access only. The IETF feels that the authentication mechanisms are inadequate for update access but has allowed the standard to proceed for read access when some other means of updating is used. (See also, Hodges[7]).

In spite of the IETF reservations this version has rapidly gained wide acceptance. All the major mail clients (Netscape, Outlook, Eudora etc.) support it and stand alone LDAP servers are available from several vendors (Novell, Netscape, Lotus/IBM, Innosoft etc.) as are X. 500 gateways (Sun, Microsoft, etc.). It includes the concept of referrals and restores some but not all of the authentication and validation mechanisms of DAP. It also includes a well defined syntax for encoding distinguished names[8], attribute values[9] and search filters[10] as strings.

2.2 *Existing technologies*

The most familiar example of directory service is the rolodex or a box of 3X5 cards. Like card files, directory servers manage smallish packets of information (a directory entry or card) associated with a named persons or organizations that can record a diverse set of attributes. Directory service is not simply a billion card rolodex however because the servers don't just maintain the information, they will search through it for you and return only selected information. Servers can also suggest other servers (referrals) to enlist in the effort, i. e., you may end up searching several directories to get a result but not need to he aware of

this.

Directory servers do not perform the join operation that relational databases use to combine information from different tables. Instead they offer increasing flexibility in representing and searching for information. An attribute of an entry in a directory may be missing or have multiple values. While it is possible to represent multiple values in relational form it requires introducing new tables and joins, i. e., substantial overhead and complexity so it is generally not done unless it is necessary. Missing values are usually supported in relational databases but usually require storing a special missing data value. The low overhead for missing and multiple values in a directory makes it much easier to accommodate rarely used attributes and occasional exceptions such as persons with multiple telephone numbers. Directories are organized and searched hierarchically. Again it is possible to do this with SQL stored procedures and temporary tables but it is awkward.

A directory in many ways is an object oriented database. The difference between directory service and a traditional OODB is that a directory associates attributes with objects but not methods and that binding to the attributes is done at runtime as a lookup operation rather than at compile time. The first means that you can retrieve arbitrary data from an object but the only functions you can perform on it are the search, add, modify, delete etc. defined by LDAP. The latter consideration is similar to the relationship of interpreted BASIC to a compiled higher level languages and with analogous benefits (to the programmer and user) of simplicity, flexibility and rapid development and costs (to the computer) in performance.

Frames are a data structure commonly used in artificial intelligence shells. Their key feature of frames is that they inherit properties from their parents. Directory entries do not do this because objectClasses inherit attributes but not attribute values from their parents. However, this functionality can easily be implemented on the client side. One simple scheme is to first look for the attribute in the named frame and if it is not present strip off the rdn and look for the attribute in the frame named by the parent dn (if it has objectClass=aiFrame). A more flexible scheme would be to define an entry of class aiFrame to include a dn valued attribute aiParentFrame and to trace that. Eventually it might be beneficial to move this to the server side either by defining an LDAP extension or by defining a new *ancestor scope* option for the search function.

Uniform Resource Locators (URL)are the internet standard for locating information.

For most protocols they are based in the Domain Name System (DNS) which identifies individual computers on the IP network. This presents problems when more than one computer offers access to the resource or the computer serving the resource changes with time.. Distinguished names avoid this problem and may be served by many computers, i. e., directory entries may be replicated or cached for reliability or performance and the responsible servers may change over time.

2.3 *Benefits of directories*

A major advantage of LDAP is the availability of LDAP servers and client toolkits. Standalone servers and LDAP to X. 500 gateways are available from several sources. LDAP client libraries are available for the C language from Univ. Michigan and Netscape and for the Java language from Sun and Netscape. Furthermore LDAP is a standard which is directly utilized by the clients and all clients should be able to talk to all servers. In contrast, SQL standardization has more to do with transportability of programmers and database schema than interoperability of databases.

The X. 500 information model is extremely flexible and search filters provide a powerful mechanism for selecting entries, at least as powerful as SQL and probably more powerful than typical OODB. The standard defines an extensibleObject which can have any attribute and some standalone LDAP implementations permit relaxed schema checking, which in effect makes any object extensible. Since an attribute value may be a distinguished name directory entries can make arbitrary references to one another, i. e., across branches of the directory hierarchy or between directories. Some LDAP and X.500 server[11] permit fine grained access control. That is to say that access controls can be placed on individual entries, whole sub trees (including the directory itself) and even individual attributes if necessary. This level of control is not available in most existing databases.

Referrals mean that one server which cannot resolve a request may refer the user to another server or servers which may be able to do so. During a search operation any referrals encountered are returned with the entries located and the user (or client) has the option of continuing the search on the servers indicated. This allows federation of directories which means that multiple LDAP/XSOO servers can present to the user a unified namespace and search results even though they are at widely separated locations and the implementations

may actually be very different.

2.4 *Java Naming and Directory Inteflace*

The Java Naming and Directory Interface[12] (JNDI) is a standard extension to the Java language introduced by Sun. It includes an abstract implementation of name construction and parsing which encompasses the X. 500 name space among others and an abstract directory that is essentially the X. 500 information and functional models. Specific implementations (service providers[13]) are available for LDAP, Network Information Server (NIS) and even the computers own file system.

JNDI removes many of the limitations of LDAP as an OODB by providing a standard way to identify the Java class corresponding to a directory entity and instantiate it at runtime. It is also possible to store serialized Java objects as attribute values. Sun has proposed a set of standard attributes and objectClasses to do this.

3 **Naming**

3. 1 X. 500 *Distinguished Names*

When represented as a string[8] (essentially always with LDAP) a distinguished name is a comma separated list of attribute value pairs and is read from right to left. If an value contains special characters such as commas it must be quoted and in any case initial and final white space around attributes or values is ignored. For example, "cn=Wayne Moore, ou=Genetics Department, o=Stanford University".

Location names have as their root (right most) component the countryName or c attribute with the value being one of the IS0 standard two letter country codes, for example c=US. Such names can be further restricted by specifying a stateOrProvinceName abbreviated st and a locality abbreviated 1, for example "l=San Francisco, st=Califomia, c=US".

Organizational names have as their root the name (registered with ISO) of a recognized organization and may be further qualified with one or more organizational units, for example "ou=Department of Genetics, ou=School of Medicine, o=Stanford University".

Domain names as used by the Domain Name Service (DNS) are represented with the dc attribute, for example, "dc=Darwin, dc=Stanford, dc=EDU".

Names of persons. There are two conventions for naming people. The older uses the commonName or cn attribute of the Person objectclass but these are not necessarily unique. Some directories use the userId or UID attribute of inetOrgPerson, which is unique. Since uniqueness is important for scientific applications the latter will be used. The remainder of a persons dn is usually either an organizational or geographic name, for example "uid=wmoore, o=Stanford University" or "cn=Wayne Moore, l=San Francisco, st=California, c=Us".

3.2 *Encapsulating and extending existing nomenclatures*

The following examples are chosen because they are referenced by the flow cytometry objects introduced below.

Gene loci, for example, "locus=Igh-1, o=Professional Society or locus=New, cn=Leonard Herzenberg, ou=Department of Genetics, ou=School of Medicine, o=Stanford University".

Gene *alleles,* for example, "allele=a, locus=Igh-1, o=Professional Society or allele=l, locus=l27, ou=Department of Genetics, o=Stanford University".

CD antigens, for example, "specificity=CD23, o=Human Leukocyte Differentiation Workshop".

Literature references in the scientific literature are have essentially achieved the benefits of distinguished names without an explicit central authority. However representing them as distinguished names will facility mechanical processing. For example, "title=" A Directory of Biological Materials", volume=1999, o=" Pacific Symposium on Biocomputing". A true directory of such literature references would be of obvious value.

3.3 New *nomenclature schema*

The following schemas arose from work on storing information about flow cytometry data in directories.

Monoclonal antibodies are distinguished by cloneName or clone which is unique within the parent entity which must be an investigator or organization.

Lymphocyte differentiation antigens, a thesaurus of the target specificities of monoclonal antibodies. Would include but not be limited to the official CD names.

FACS instruments are distinguished by the cytometer attribute which must be unique with respect to the organization parent, for example, 'cytometer=Flasher II, ou=Shared FACS Facility, o=Stanford University".

FACS experiments are distinguished by the protocolIdentifier or protocol attribute which must be unique with respect to the parent which may be a person, and instrument or and organization or some combination, e. g., "protocol=1234, cytometer=Flasher, uid=Moore, ou=Shared FACS Facility, o=Stanford University".

FACS samples are distinguished by a unique protocolCoordinate which must be unique within the parent FACS experiment, e. g., "coord=A12a, protocol=12345, cytometer=Mollusk, ou=Shared FACS Facility, o=Stanford University".

4 Biological Object Schema

X. 500 defines a sparse set of standard types and standard objects mostly for describing persons and documents and more suitable for business than scientific use. However if types were added for scientific use, particularly real numbers and possibly dimensional units, much scientifically relevant information could be conveniently stored in and accessed from directories. The following minimal set of objects for the field of flow cytometry is presented to lend concreteness to the discussion. A fuller and formal definition will follow.

Table 1: Scientific Investigator

objectClass	cis	ScientificInvestigator, InetOrgPerson, organizationalPerson, person
UID	cis	User identifier must be unique in context
ou	cis	From distinguished name
o	cis	From distinguished name
professionalName	cis	Author name(s) used in the literature
professionalSpeciality	cis	For example "Cellular Immunology"
professionalAffiliation	cis	For example, "National Academy of Sciences"
professionalPublication	dn	ScientificPublication of which this is an author.

Table 2: Scientific Instrument

objectClass	cis	ScientificInstument
cn	cis	Common name
ou	cis	From distinguished name
o	cis	From distinguished name
instrumentManufacturer	dn	For example, ou=immunocytometry Systems, o=Bectorn Dickenson
instrumentModel	cis	For example, "FACS-II"
instrumentSerialNumber	cis	Manufactures id
responsiblePerson	dn	Dn of a person responsible for the instrument

Table 3: Scientific Publication

objectClass	cis	ScientificPublication, document
title	cis	Title
volume	cis	Volume
ou	cis	From distinguished name
o	cis	From distinguished name
pages	cis	Range of pages
reference	dn	Distinguished name of publication referenced by this publication
citation	dn	Distinguished name of a publication which referenced this one
author	dn	Distinguished name of author

Table 4: Monoclonal antibodies

objectClass	cis	MonoclonalAntibody
clone	cis	Unique clone name
o	cis	From distinguished name
ou	cis	May be part of distinguished name
UID	cis	May be part of distinguished name
cn	cis	Common name(s)
specificity	dn	Distinguished name of specificity
creatorDn	dn	Distinguished name of person or organization that created the clone.
titre	float	
concentration	float	
manufacturer	dn	Designated name of manufacturer
heavyChain	dn	dn of heavy chain locus or allele
lightChain	dn	dn of light chain locus or allele

Table 5: FACS instrument

objectClass	cis	FlowCytometer, scientificInstrument
cn	cis	Common name
instrumentManufacturer	dn	For example "ou=Immunocytometry Systems, o=Becton Dickenson"
intrumentModel	cis	For example, "FACS-II"
instrumentSerialNumber	cis	Manufactures identifier

Table 6: FACS experiments

protocolIdentifier	cis	Uniquely identifies protocol in context
UID	cis	May be part of distinguished name
instrument	cis	May be part of distinguished name
o	cis	May be part of distinguished name
ou	cis	May be part of distinguished name
instrumentDn	dn	Distinguished name of a scientific instrument
archiveURL	url	URLs of archive file corresponding to this experiment
dateCollected	date	
numberOfSamples	int	Number of samples collected

Table 7: FACS sample

protocolCoordinate	cis	Uniquely identifies sample in protocol
protocolIdentifier	cis	Uniquely identifies protocol in context
UID	cis	May be par of distinguished name
instrument	cis	May be part of distinguished name
o	cis	May be part of distinguished name
ou	cis	May be part of distinguished name
cn	cis	Common name
title	cis	Experiment title
description	cis	Description of the sample
sampleLabel	cis	Label for the sample from the protocol
investigatorDn	dn	Distinguished name of the investigator responsible for collecting the data
instrumentDn	dn	Distinguished name of a scientific instrument
dateCollected	date	
startTime	time	
endTime	time	
numberOfMeasurements	int	Number of components measured for each event
numberOfEvents	int	Number of events in the sample
URL	url	URLs of date file corresponding to this sample

5 Conclusion

This paper examines the problem of computer-assisted communications in flow cytometry in particular, and biology in general, from the point of view of the emerging standards for computerized directory service. Following Schulze Kremer[14]: "To improve the current situation of non-unified and ambiguous vocabulary, the only solution is to develop a core of commonly agreeable definitions, and using these, to implement user interfaces to and between databases". As an example of how this goal can be accomplished, I have outlined how X. 500 directory services accessed via LDAP from lightweight clients can be used to create and manage a unique namespace in the flow cytometry domain.

We plan to produce a concrete and useful implementation of a directory of the FACS experiments and sample data collected at Stanford, the National Institutes of Health, Fox Chase Cancer Center and the University of Iowa. We also plan to create a registry of

monoclonal antibodies based on input from the manufacturers and other interested parties such as the Human Leukocyte Differentiation Workshop. This work will be proposed for standardization to the National Information Standards Organization (NISO), a non profit organization accredited by the American National Standards Institute (ANSI) for information standards development, or to the working group on Accessing, Searching and Indexing Directories (ASID) of the Internet Engineering Task Force (IETF), which is responsible for intemet standards activity.

The wide use and importance of flow cytometry in basic and clinical science today means that our directory will rapidly become a significant resource for the field. In addition, this project will make the primary data from flow cytometry and monoclonal antibody production available to the wider biomedical community, as is already done for gene sequence data. We believe that there are many other fields and instrument methodologies for which this would he a great benefit.

Acknowledgments

This work was supported by grants NLMO4836 from the National Library of Medicine and CA42509 from the National Cancer Institute, both in the National Institutes of Health. .

References

1. D. R. Parks, "Flow Cytometry Instrumentation and Measurements" in Handbook of Experimental Immunology, Blackwell Scientific, (1996)
2. V. T. Oi, P. P. Jones, J. Goding and L. A. Herzenberg, Advances in Microbial. (1978)
3. "Data Communications Networks Directory", Recommendations X. 500-X. 52 1, Volume VIII, IXth Plenary Assembly, Melbourne, (CCITT, 1988)
4. T. A. Howes, "The Lightweight Directory Access Protocol: X. 500 Lite", CITI Technical Report 95-8 (1995)
5. W. Yeong, T. Howes, and S. Kille, "Lightweight Directory Access Protocol," RFC 1777, (1995)
6. M. Wahl, T. Howes, S. Kille "Lightweight Directory Access Protocol (v3)." RFC 2251 (1997)
7. J. Hodges, "An LDAP Roadmap &FAQ", Kings Mountain Systems, (1998), URL http://www. kingsmountain.com/LdapRoadmap.shtml
8. S. Kille, "A String Representation of Distinguished Names", RFC 1779, (1995)

9. Wahl, M., Coulheck, A., Howes, T., and S. Kille, "Lightweight Directory Access Protocol (v3): Attribute Syntax Definitions", RFC 2252, (1997).
10. T. Howes "The String Representation of LDAP Search Filters." RFC 2254 (1997)
11. "Netscape Directory Server Administrators Guide", Netscape, (1997)
12. "JNDI: Java Naming and Directory Interface", Sun Microsystems, (1998)
13. "JNDI SPI: Java Naming and Directory, Service Provider Interface", Sun Microsystems, (1998)
14. S. Schulze-Kremer; "Ontologies for Molecular Biology" Pacific Symposium on Biocomputing 3: 693-704 (1998).

Protocol Editor Specification (Synopsis)

Reagents

This feature allows the user to specify the reagents used in a protocol and also to summarizes visually the reagent list for the user. In the demo it is the upper left panel on the main screen. There are two parts, the "palette" which is a list of individual reagents and the "cocktails" which represent combinations of reagents which occur frequently.

Reagent Attributes

Distinguished Name

Clone ID

Common Name

Specificity

Titre

Fluorochrome Name (&Lot #if neiessary)

Concentration (optional)

Reagent Palette

The reagent palette is populated by copying or referencing entries from a number of sources.

Reagent Catalog

Current intention is for Stanford to implement the database this via JNDI. There will need to be some browser and search functions

"Bag of Tricks"

The user may have a "bag of tricks" which is a light-weight db of reagents probably serialized into a local tile in which they store frequently used reagents and copies of cataloged reagents for use on portables etc.

Other protocols

The user may open multiple protocols and copy/paste or drag/drop reagents between protocols.

New manually entered reagents

The user may define a new reagent by supplying the required information. An attempt should be made to check for conflicts with the catalog and in most cases to try to catalog the new reagent.

Reagent Sets

Often there are groups of reagents which are used together repeatedly in a protocol. The demo uses a tree widget to visualize this where the nodes are "cocktails" and the leaves are reagents. This is good for some experiments but questionable for others particularly high numbers of colors. A grid with "colors" as columns and "cocktails" as rows with individual reagents in the cells might work better for some purposes. We may need to do some explicit prototyping before finalizing this. You should be able to copy whole reagent sets and have the individual reagents merged correctly into the palette.

Consistency checking

It is rare that more than one reagent of a given color is used in a cocktail or crossing experiment but not impossible.

The user should be warned of consistency violations but allowed to enter them.

It should also be able to be told to accept them without comment in the future (for this protocol).

Two step stains 7

Samples

This feature allows the user to define the cell samples which will be stained. In the demo it is the lower left panel and it is implemented as a grid widget.

Sample Attributes

Each column in the table is a sample attribute. It has a string name and a data type which ultimately should be from the same set that JMP uses (or a super set) but string and number would get us started.

The user can define new attributes, redefine an existing attribute or copy one or

more from other protocols or possibly from the bag of tricks.

The user can reorder the columns at will by dragging.

Sample Palette

Each row in the palette represents one sample

You must be able to copy selected groups of samples to the experiment models.

You should be able to copy rows and groups of rows between protocols (and perhaps the bag of tricks).

Create rows, duplicate one or more rows, delete rows, insert unique #, fill columns or ranges in columns

Collator

The collator allows the user to sort and resort the sample palette as needed and also facilitates logical group selection. It *is* implemented in the demo by a special "colomn" which has an icon label no data and a different background.

The user can drag the collator like any other column and can drag other columns over it.

The data grid is always sorted by the attributes to the left of the collator in left to right precedence.

The background of the cells in the first column are colored with two colors and the color toggles every time the value of that column changes. The second column changes color every time the value in either the first or second column changes and similarly for the others, i. e., in each column a block of color represents all the rows which match at and to the left of the column. To the right of the collator the background follows the pattern of the right most sort column.

Selecting any cell in a column to the left of the collator means selecting all the rows which match in this column and to the left, i. e., the complete color block. It can be copied as such to the experiment modes.

Ideally this should be a logical definition, i. e., if new entries are made which match the criterion they should be propagated forward but this may be too complicated.

Symbolic Selection?

Sample volume, cells, concentration?

Experiment Models

Experiment models are stereotyped fragments of protocols which may occur repeatedly in an actual protocol. By far the most important is a "crossing experiment" followed by "compensation controls", the others might be Vl .l of V2 features.

Crossing Model

Many experiments are essentially combinatorial, i. e., this set of reagents or reagent cocktails applied to each sample in a group of samples. Typically it will correspond to some N X M grid of wells in the staining plate. An experiment might have 1 or >lO of these. Very large experiments are probably better off with the screening model.

The experiment model widget has two side by side panels and you can copy reagents or reagent cocktails into one and samples into the other.

You can also copy or move reagents and samples between models within one protocol. Copying between protocols without first copying the sample/reagent info is probably too complicated.

You can delete entries from a model but they remain part of the sample palette even when there are no references. You cannot delete a row from the sample column while it is referenced by a model (or the combined delete must be OK'd).

The user may change the staining volume (typically 100~1)for the model. The default is a user preference.

The user can transpose the layout of the model, i. e., N X M vs M X N grid pattern on the plate, the default being a user preference.

Titration Model

The user specifies a target sample and a reagent and then a range of dilutions 2,4, 8.. . or 10, 20, 50, 100 being typical.

Layout as a single column (or row)on the plate.

Screening Model

The user specifies a reagent cocktail and a large number of samples which are quasi-automatically named.

FIuorescence Compensation Controls

For each dye (or dye lot)which occurs in an experiment model allow the user to specify a sample to be used as a control. Usually it will be one of the samples which is stained with the reagent.

Unstained Controls

For each sample defied an unstained control

Plate Layout

Staining is typically done in "micro-titre" plates which are an 8 X 12 array of small wells. Other form factors should be available however probably as a user preference. An experiment may require several plates (all of the same form factor). Some users prefer to skip every other row and/or column.

In the demo this is the far right panel. The user selects a model and by drag and drop places the corresponding N X M grid onto the plates. It may need to cross plate boundaries. When the user selects a well in the plate map the corresponding experiment model should be scrolled on screen and highlighted and the sample and reagent cocktail information should be highlighted both in the model and in the palettes.

Do we need finer positioning controls?

Printed Documents

Protocol Worksheet

For each sample add up the total volume (cell number?)needed.

For each reagent used add up the total volume used in the experiment.

For each cocktail, a worksheet showing how much of each reagent to mix into each cocktail to make enough for all samples at the appropriate dilution. If the concentration is known and the user prefers it compute and display the actual concentrations.

Pipetting Map

Currently the desk protocol editor prints an image of the plates for each reagent/sample with the well which receive that item marked. The demo doesn't do it at all. Need more feedback on how useful this is and in what form.

User Preferences

Samples =rows X Reagents =columns or transposed

Default well total volume and sample volume (cells?).

Plate/rack form factor (8 X 12 by default).

Layout rules for example skip columns etc.

Bag of tricks containing commonly used reagents and sample models. Stored locally.

Does the *user* wish to see absolute concentrations when available.

Reagent databases to search (search order preference?)

Default titration series?

Menus

File

Edit

Models

Help

Data Collection

We have not adequately specified how this will interact with setup, calibration and data collection. The desk model is that the user selects a well on the plate map, a dialog with the annotation information for that well is presented and the user can edit the info. (it's not clear what that means in this case since it may come from a database). The user can start, pause, abort or finish collection. Starting collection should start the clock for kinetics data. One likely scenario is that the protocol editor will then make an entry describing the sample using JNDI

For setup it needs to export a list of the fluorochromes used so a suitable setups can be identified.

Research Plan

A. **Specific Aims**

In this and the accompanying Phase I SBIR Proposal, we propose to develop an Internet based Research Support System (IBRSS)that will support the integrated processing and interpretation of large data sets acquired with modern biomedical instrumentation. To this end, we propose to develop and deploy a highly innovative Internet infrastructure with specific tools designed to support the collection, annotation, storage, management, retrieval, analysis, and sharing of data from flow cytometry, DNA micro array and other data-intensive biomedical instruments. This system, which will facilitate access to past and present data, will provide much-needed permanent recording capability for patent and other purposes, including the possibility of recovering data from studies "even after the postdoc has left the lab."

In Phase I, we proposed to build the core IBRSS system that will provide the central computing capabilities necessary to receive and catalog data annotation information acquired at remote sites and to move large instrument-generated data sets from remote sites to the central IBRSS site. In addition, we proposed to enable remote users to search the catalog and retrieve data stored in the system. In Phase II, we propose to complete this system by providing tools for acquiring study and experiment annotation information (protocols), by improving the catalog searching tools to enable searches on additional information, and by implementing tools that will enable launch of third party analysis and viewing software and other tools to facilitate data usage and interpretation.

To accomplish the above, we plan to achieve the following Specific Aims in Phase II:

1. **Create interfaces to capture annotation information (study and experiment descriptions)**

a) **Study protocols** capture the hypotheses to be tested and the factors that go into them, including subjects, treatments, experiments and the timeline for an overall study

b) **Experiment protocols** acquire annotation information to define the subset of subjects for which data will be collected, the set of samples to

be obtained from the subjects, and the analytic procedures and data collection instruments used to analyze the samples

c) **Sample-treatment** protocols acquire annotation information to define the subdivision (aliquotting)and the treatment (reagents and conditions)for a set of samples for which data will be collected by a single analytic method (usually a single instrument).

2. **Develop novel methods for automatic aspects of protocol specification**

a) Capturing the model

b) Automating definition of plate/rack layouts and creation of reagent "cocktails"

c) Checking for inadvertent omission of controls and automatically supplying controls

d) Facilitating data collection and analysis

3. **Co-operate with instrumentation manufacturers to develop data collection modules that utilize the protocol information captured by ScienceXchange (no support requested)**

4. **Create tools to enable launch and use of third-party analysis and visualization programs**

5. **Establish test sites**

a)Maintain the current alpha test site at Stanford University

b)Establish two beta test sites at Phase II start

i)Fox Chase Cancer Center

ii)University of Iowa School of Medicine

c)Establish additional beta test sites later in Phase II

i)Multi-center clinical cancer research consortium

ii)European and Japanese research sites

B. **Background and Significance**

Research involves many types of data collections -text, image, graphics, video, voice and numeric coming from many sources and drawn together for analysis, interpretation and reporting of final results. However, researchers needs are much the same whether their studies focus on genetics, flow cytometry, immunology, neurobiology, plant pathology, oceanography or human disease treatment and clinical trials. Therefore, in developing the technologies outlined here, we will work towards meeting these needs and creating an ideal working environment for the biomedical researcher -a low cost, easily accessible

workspace that offers tools to help gather, manage, store, analyze and interpret raw and annotated data.

The basic motivation for, and overall description of, the IBRSS project is described in the *Background and* Significance Section of our Phase I application. To facilitate reviewers'access to this material, we have reproduced key parts of the application in the indented text section that follows:

Although there are wide variety of tools that purport to help scientists deal with the complex data collected in today's laboratories, virtually all of these so-called Laboratory Information Systems (LIMS)or Electronic Laboratory Notebook systems (ELNs)approach data collection and management from the perspective of final data output and interpretation. To our knowledge, none of these systems addresses the basic needs of the bench scientist, who lacks even minimal tools for automating the collection and storage of data annotated with sufficient information to enable its analysis and interpretation as a study proceeds.

The absence of automated support for this basic laboratory function, particularly when data is collected with today's complex data-intensive instrumentation, constitutes a significant block to creative and cost-effective research. Except in very rare instances, the study and experiment descriptions scientists need to interpret the digitized data these instruments generate are stored in paper-bound notebooks or unstructured computer files whose connection to the data must be manually established and maintained. The volatility of these connections, aggravated by turnover in laboratory personnel, makes it necessary to complete the interpretation of digitized data as rapidly as possible and seriously shortens the half-life of data that could otherwise be mined repeatedly.

In addition, because notebook information is difficult to make available to other investigators, particularly at different sites or across time, laboratories that would like to make their primary data available to collaborators or other interested parties are unable to do so. Thus, although computer use now facilitates many aspects of research, and although the Internet now makes data sharing and cooperative research possible, researchers are prevented from taking full advantage of these tools by the lack of appropriately tailored computer support for their work.

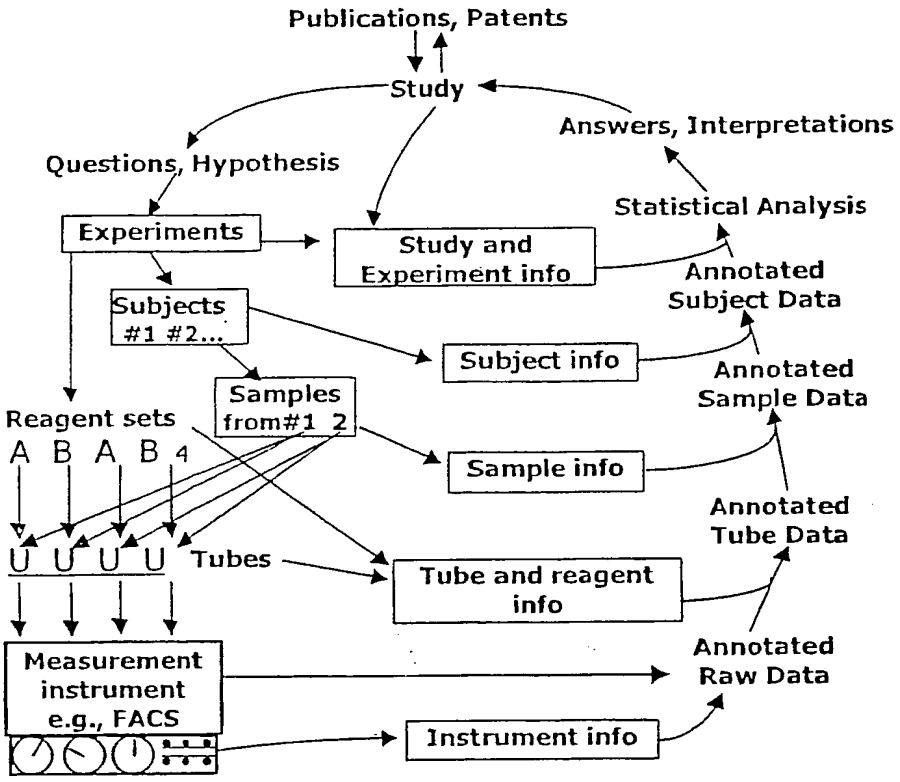
Finally, because what computerized support for research currently exists has developed piecemeal, usually in response to needs encountered during collection of particular kinds of data, no support currently exists for providing lateral support to integrate different types of data collected within an overall study. For example, although

automated methods for collecting, maintaining and using DNA micro array data are now becoming quite sophisticated, the integration of these data with information about the source of the material analyzed or with results from other types analyses done with the same material is largely a manual task requiring recovery of data and information in diverse files at diverse locations known, often, only to one or a small number of researchers directly concerned with the details of the project. In fact, it is not uncommon for individual bench scientists to be forced to repeat experiments because key information or data was "misplaced" or its location lost over time....

Protocol editors to acquire annotation information. Managing data - storing it, finding it and, most importantly, extracting answers to the questions the experiment was meant to answer -requires the ability to combine the data itself with annotation information recorded in study and experiment protocols, which are constructed by the investigator prior to data collection and dictate how the experiment will be done. The complexity of this process is outlined in figure 1, which charts the way information is collected, analyzed and stored in typical biomedical studies.

In laboratory parlance, a series of data collections usually constitutes an experiment and a series of experiments usually constitute a study. Notes are made, usually on paper but more often now in computer files, to define the overall study plan and, as experiments are added, to describe the subset of subjects in each and the experimental manipulations to be performed with the subjects or with cells or other material acquired from the subjects. Finally, notes relevant to the data collection, and the data itself, are added to the collection of experiment notes. In general, studies take months or years and individual experiments may take days, weeks or months to complete. Therefore, notes containing annotation information needed to support data interpretation may be spread throughout a journal-style notebook or collected in a variety of directories and files and must be reassembled before the full results of a study can be assessed.

Figure 1: Information flow in a typical biomedical study



Most researchers are not conscious of the various steps in this process. They treat and teach them collectively as the age-old art of experimentation and firmly believe that nothing can be done to improve the way notebook entries have to be handled. Consequently, instrument manufacturers are seldom urged to provide support for more than just collection of the data and initial calculations to convert it to usable units. However, the structured collection of annotation information should not be neglected. Although this is clearly a difficult problem to attack, the expanded research capabilities that would result from its solution merit the effort involved in finding ways to acquire annotation information at the planning stages of studies and experiments so that the information can be used later in the process to interpret the data that is collected.

The basic IBRSS protocol elements. Analyzed from a systems point of view, the capture of information required to utilize machine-generated data in a typical experiment is conceptually organized into several information capture protocols: 1)*study protocols,* which capture the hypotheses to be tested and the factors that go into them, including

subjects, treatments, experiments and the timeline for an overall study; 2)experiment protocols, which dictate the details of treatments for samples in the experiment; 3)*data collection protocols,* which specify the samples and reagents that will be put in the test tubes, the planned incubation time and conditions, the specific instruments that will be used for data collection and any instrumentation settings unique to the experiment. These protocols also contain global identifiers for reagents and appended notes concerning anomalies that occurred during sample addition, incubation or data collection; and, 4)analysis *protocols,* which specify calls for analyses, e. g., to determine subset frequencies, median fluorescences, etc. These protocols may be specified before and/or after data collection and will likely be passed to co-operating third-party analysis software.

As indicated above, researchers currently enter protocol information in relatively unstructured fashion into their notebooks or into computer files that provide aspects of notebook function. However, once the underlying data structures are identified, developing automated methods to enable structured collection and storage of this information become quite feasible. In essence, the task of collecting the annotation data relevant to individual experiments and studies resolves to developing user interfaces that encourage and facilitate capture of the information specified in each of several protocols....

Developing a well-indexed system that will build and maintain a permanent linkage between context information and primary data has several benefits over and above the support it provides for bench scientists. Professors, for example, would be able to find data after the student or post-dot who collected it has left the laboratory. Furthermore, research managers will be able to directly access their teams'research product rather than depending on staff to maintain and provide file locations for data. In addition, data required for patent-related activities will be readily accessible even years after it was collected, independent of whether the researchers who collected it are still available and/or able to recover the required items.

Creating an Internet-based system with these characteristics will also solve one of the central documentation problems in modern research. The existence of a stable, Internet-accessible data archive readily permits published studies to refer to primary source data and also provides facilities for using the source data to find published studies in which it is referenced. Thus, this system provides the infrastructure necessary for creation of an Internet-based international repository

(either central or distributed)for the primary data from which published data is derived.

Similarly, in the medical arena, the development of this system will enable creation of a webbased system for cataloging and providing access to standardized data sets from flow cytometry, DNA microarray and other instrumentation used to diagnose and monitor treatment of a wide variety of diseases. Thus, it would provide the infrastructure for a much-needed international repository of biomedical information to provide universal access to critical information that currently resides in only a few, well-financed medical institutions.

In sum, there is a growing recognition of the need for services that remove barriers to effective research. Research activities in academic, government and private sector laboratories are all restricted by the lack of automated support for recording, archiving and accessing data. To meet this need, and to make services available to laboratories that cannot support sophisticated local computer technology, we plan to create and field an integrated, Internet-accessible system capable of supporting the secure collection, annotation, storage, management, cataloging, retrieval, analysis, interpretation and sharing of data from flow cytometry, DNA microarray, imaging and other data-intensive biomedical instruments.

Specification of the modules in this system clearly requires a specialized interaction between software -engineers and working biologists and medical scientists. The FAGS/Desk system in the Herzenberg laboratory was evolved as just such a partnership. Therefore, although outdated, the principles that underlie its basic design are highly informative with respect to our goals here. However, to develop a broader set of annotation capture tools requires interaction with scientists in more than one laboratory and using multiple biomedical technology. Thus, our plan here is to rapidly add three beta test sites to the Stanford alpha s'ite and to craft our annotation tools and other user interfaces to meet the needs of scientists working in this broader arena.

We still plan to center our initial design work around the development of support for flow cytometry (FACS)work since this technology is central to the work of researchers at the two initial beta sites have legacy FAGS/Desk archives that IBRSS can immediately import. However, we will move as rapidly as possible to include other technologies, since these are also central to the work of the researchers at all sites. To this end, we have

already made a first pass at designing LDAP structures that can be used for genetic studies (see Appendix XX).

D. **Experimental Design &Methods**

Background to Methods:

The Phase II funding requested here will support the development of designs and prototypes for user interfaces (protocol builders)that will enable structured capture of detailed experiment and study descriptions and will provide unique support for experiment planning and data analysis. These interfaces will be designed to function in a JAVA-based client-server environment (IBRSS)and will support the structured entry of study and experiment information by presenting the user with a broad array of relevant *standardized-choices* to be selected for entry into a *standardized* set of fields. Where useful, we will design the interfaces with associated "Wizards" to guide the user in making choices among options and in structuring the protocol so that it contains controls appropriate to the experiment.

As indicated above, we will begin by creating a protocol builder for FACS studies, since 1)there is longterm experience with the "protocol builder" in FAGS/Desk, which has been used to capture the annotation information (albeit minimal)that will be transferred to IBRSS during Phase I of this project; 2)users at all of the test sites have experience with using the FAGS/Desk protocol builder; and, 3)the richness of the kinds of data collected with FACS instruments offers a demanding model fordeveloping protocol builders to support data and annotation information from data-intensive scientific instruments. Similar protocol builders, designed to serve other DNA microarray, imaging and other biomedical technologies, will be developed using this overall model.

The FACS itself simply measures cell-associated fluorescence and light scatter for individual cells passing single file, in a laminar flow stream, past a set of light detectors. The cell-associated fluorescence results from "staining" (incubating)cells with fluorochrome-coupled monoclonal antibodies or other fluorogenic or fluorescent molecules that bind specifically to molecules on, or in, cells. As each cell passes the FACS detectors, it is illuminated by a set of lasers that excite the fluorescent molecules associated with the cell. This causes the cell to scatter light and to emit fluorescent light at wavelengths defined by the associated fluorochromes. The amount of light derived from the cell is then measured by the detectors, which are set to measure the light emitted at particular wavelengths or scattered at particular angles.

The measurements made by each of the FACS detectors are processed, digitized,

joined and recorded on a cell-by-cell basis in a data file that has one such record for each cell analyzed. For a sample stained with a given set of reagents, 4-13 measurements per cell (depending on the FACS instrument)are collected for at least ten thousand, and sometimes up to 5 million cells. This "FACS analysis" usually takes less than a minute and lo-100 samples are typically passed through the FACS in a single session.

Before collecting FACS data, FAGS/Desk users typically file a protocol in which they enter short free-text descriptions of the reagents and cell types used in each sample. This information is displayed during data collection and permanently associated with the data once collected. It is then maintained within FACSjDesk until the user calls for it to be exported, along with the actual data, to analysis/visualization modules. Cooperating analysis modules (e. g., FAGS/Desk itself or in the FlowJo software)use this information to label axes on graphs and column heads on tables; IBRSS will use it additionally to catalog the data so that it can be retrieved based on any combination of information included in the protocol.

The new protocol builders will collect standardized, rather than free-text, entries wherever -possible to make catalog searching more efficient. In addition, they will have modern interfaces (rather than the antique interface in FAGS/Desk)and associated Wizards. Thus, as the protocol builders mature and are modified according to user feedback, they will constitute excellent models for the development of protocol builders to serve biomedical instrumentation other than FACS.

**Start Confidential

Definitions

Experiment model vs data model. Programmers working with complex systems and databases commonly begin by creating a data model and working from that. However, for our purposes here, a data model per se is likely to be too concrete. The following five data model items loosely define a somewhat abstract approach to the experiment model: the first two items give an example of the abstract data model; the next three provide concrete examples of samples that are encountered in the protocol editor. We use FACS studies here as concrete examples. Similar definitions, tailored to other technologies, will be developed as our work proceeds.

Attributes. In statistics-speak, an attribute is called a "random variable", but this terminology seems only to confuse biologists. Attributes have names (usually unique within a restricted framework). An attribute's values may be "nominal", "ordinal" or "continuous". It may be an "independent" or "dependent" variable in a model. These are

hints to the statistics assistant as to how to treat this attribute as a factor in a model.

An attribute may be "internal", "external" or "computed." An internal attribute is created by the protocol editor and stored in the experiment document. External attributes are links to data external files or databases, e. g., JMP tables, SQL databases, or LDAP directories that must be keyed by some attributes of the sample. Some examples in databases in a clinical study include demographic data, vital signs or clinical incidents. Computed attributes are scalar-valued statistics computed from the cell data for a FACS Sample. For completeness, one could make a case for attributes computed from the existing attributes of a sample as well.

Databases entries mapping subject id's to patient information and assigning trial arm (i. e., drug vs. placebo)are special attributes and must be isolated and protected specially. External or computed attributes might be cached for efficiency, but private or blinded information should not be cached,

Abstract Sample. This is a placeholder that defines things common to the concrete samples defined below. Most importantly, a sample may associate one or more attributes with values (of the appropriate type)and inherits attributes and their values from a super-sample if it has one. A sample may be excluded; if it is, all of its sub-samples are excluded as well. Who excluded the sample, and when and why it was excluded should be part of the record. Examples of an exclusion might be a non-compliant patient, a blood draw which was bad, or an instrument malfunction such as a nozzle clog. Excluded samples may be graphed and analyzed using FACS-specific methods but are normally excluded from meta analysis, i. e., are not exported to JMP etc. for final experiment or study analysis. Obviously one way of handling the exclusions is as a special form of attribute.

Study or experiment subjects. The value assigned to a subject is essentially an identifier that is unique (at least)to an experiment and may be unique with respect to the study of which the experiment is a part. In addition, it may even be unique globatly (e. g., a distinguished name). A subject is technically, that is statistically, a sample from a larger population (say of mice or men). It may have one or more attributes and may have or require an attribute, e. g., "subject type" for "human", "mouse", "cell line," etc. but it should be a hint to the "protocol expert" on how to initialize the default and predefined attributes at the interface level, not a polymorphism in the data model.

The identifier (or identifiers)must allow for linking the data to external sources but definitely should not include identifying information about human patients. Inclusion of such information would subject experiment data to stringent legal requirements for access

control and encryption that would interfere with collaboration. This identifier may also be used as a key for blinded data, which is defined in the study data model but not available until after FACS analysis (and the rest of the data collection)is complete.

Cell Sample*. Cell samples are obtained from subjects at a particular time. Subjects may be sampled more than once, either by taking multiple samples at a single sitting or (usually)by sampling repeatedly -over time. Cell samples inherit the attributes of the subject and may add new ones including: time of sampling; the sampled tissue (e. g., peripheral blood, bone marrow); how the sample was handled (e. g., ACD, Heparin, put through Ficoll, etc); and the sample's role in the study (Screening, 0 week, 2 week, . . . 8 week). Cell samples may need to be able to be linked with external data such as vital signs or clinical lab reports, for example, to compute absolute CD4 counts.

Stained Sample. Cell samples are usually divided (aliquotted)and treated with different combinations of reagents. A stained sample must be identified in terms of an element in a specific experiment model, e. g., a specific well or test tube in a cell sample by reagent cocktail crossing. In addition to the attributes of the cell sample, a stained sample copies all the attributes which are factors in the staining model. For each color of reagent in the reagent cocktail, a new nominal attribute named for the color is added whose value is typically the specificity of the first antibody in the reagent complex of that color. This is used later in labeling visualizations of the sample. For supporting the bench work, a stained sample has a target cell count and target volume that are needed to compute the pipetting instructions. It must be associated with some coordinate that allows it to be identified to the data collector (currently simply row and column).

A technical assistant for an experiment in which samples are to be stained with several sets of reagents, each in a separate test tube, can advise the user as to the minimum number of cells required per sample for the sample to be aliquotted into all of the specified staining tubes.

For mouse experiments, the subjects will be identified by strain and either by animal number or cage number. Basically, the user will create a list subjects that includes mouse *strain* and a model that, for example, has crossed attributes such as *immunization, treatment* or *mouse strain* that represent the actions and variables in the experiment. The user can then assign subjects to the groups defined by the data model or they can request the assistant to distribute the subjects. Since *mouse strain* appears in both models, the assistant must account for this in making the assignment. The list of cell samples is then the sample model for the protocol. However, the system will remember that the cross of

immunization and *treatment* and *mouse strain* is a sub model. Everything up to this point has involved independent variables.

For a clinical trial, the data model is important enough and complex enough to warrant explicit definition, perhaps as part of a study. For example, in a recent clinical trial at Stanford, patients are identified by an anonymous identifier (nominal)with private and blinded information stored separately. Patients are also stratified into CD4 low and high (nominal)and then divided into glutathione low and high by the median FACS staining value for this parameter within each class computed separately (ordinal). Clinical lab results come in as dBase11 file (keyed by patients initials and date). Demographics and vital signs are in commonly in FileMaker databases. Patients are randomized into clinical trial arms (drug vs. placebo)by a third party. Blood is collected at 2-week intervals from 0 to 8 weeks.

The sample model for the study is *subject* crossed with *week of visit* (ordinal)and CD4 *stratum* (nominal) and then nested with *glutathione,* which is ordinal. The sub-model, which will be analyzed statistically, is week of visit crossed with trial arm and CD4 level and then nested with *glutathione* level. For a specific instance of the study protocol with a particular set of patient blood samples, the sample model will be a list of cell samples. At a minimum, the attributes of these samples will include the *patient id* and the *week of visit* and will represent an instance of the subject crossed with visit sub-model of the study. Everything so far is again independent variables.

The user must also prepare (in unspecified fashion)a reagent model of similar structure and possible complexity. For the mouse experiment, there are likely to be a small number of reagent cocktails. For the clinical trial, there were 8 or 10 cocktails (reagent sets). In either case, the reagent model is a single attribute whose value is the name of the reagent cocktail. Reagent attributes are independent. Reagent models may also in rare cases be nested, e. g., an isotype experiment performed with allotype reagents. The reagent model is crossed with a sample model (a list of cell samples)in the experiment model to generate a set of *Stained samples.*

FACS Sample. *FACS sample* is defined as the running of a *stained sampie* on a FACS instrument under _a specific set of conditions. *FACS sample* inherits attributes from the *Stained sample* and adds scale information and a start and stop timestamp (locators in the instrument log that allow reconstruction of the instrument state at the time of sampling). Analogous to sub well in the data model, if a *Stained sample* is FACS sampled more than once, each sampling is treated as a separate *FACS sample* and given a unique sequence identifier.

FACS Data Set. A FACS data value for each parameter for each cell in the *FACS* sample is collected. The values for all cells comprise a *FACS data set* for a given *FACS sample.* In addition to the raw cell data, the *FACS data set* must also include information about the scaling of the cell data. It inherits from the *Stained sample* attributes, which are used to label the data output graphically, usually the specificity for each color. FACS Data may have computed attributes which make statistical summary information about the cell sample available as an attribute of the data set

FACS Data Subset. Sometimes a FACS data subset is divided into several pieces, each containing a subset of the cells and the values recorded for those cells. These subsets inherit attributes and may also get a new independent nominal or ordinal attribute in the process. They subsets are treated as samples in their own right and thus may have computed attributes and be subject to meta analysis independently of the total sample.

Work plan

1. **Create interfaces to capture annotation information (study and experiment descriptions)**

The "protocol builders" capture the annotation information necessary to manage data from studies and experiments. During the execution of experiments, this information is initially used to identify the contents of samples during data collection. Next, it is used to retrieve data for analysis and to label analysis output (axes and column heads)with the sample and reagent information necessary for visualizing, interpreting and summarizing results. Finally, it is used to coalesce results from the individual data collections into the results of an experiment, and to coalesce the results of a series of experiments into the findings of a study. Since this crucial interpretive work may occur weeks, months or even years after all data collection for a study is complete, the strength of the annotation and data storage system that supports data collection is critically important to both the quality and the efficiency of scientific studies.

We plan to create three basic user interfaces that will acquire annotation information from users and process and transfer the information to the LDAP store. These interfaces, which will respectively collect annotation information for the *study, experiment* and *sample treatment* protocols, will be constructed as standard JAVA-based GUIs with all of the typical GUI commands (new, *save, copy, exit, etc).* Choices for annotation information for users will be offered as pull down menus, or in some cases, radio buttons. The choices will be offered as lists characteristics for a particular field, e. g., the list of choices for the SPECIES field will include mouse, rat, human, chicken, etc; the list for the

CELL TYPE field will include lymphocyte, astrodendrocyte, etc.

In some instance, users will be given the opportunity to type entries; however, this will be avoided wherever possible. Instead, an administrative function will be provided for adding missing items to lists. Administrators, who should have domain knowledge for the research group being supported, will be charged with avoiding duplications on the lists. ScienceXchange will attempt to maintain list homogeneity by augmenting centrally-supplied lists with entries defined by administrators; however, in Phase II and beyond this may become quite difficult and require application of ontological methods to resolve synonyms. If so, we are prepared to bring in consultants skilled in this area.

In general, identifiers for the kinds of protocol fields one finds in FACS (and most other)experiments are generic. However, certain items will be unique to particular studies and will have to be entered directly by users. These items will be entered only once, at the appropriate level, and will be supplied as lists thereafter. For example, *study protocol* will allow users to enter subject identifiers, which will then appear as selection lists for the *experiment protocoi generator.* Users will choose from this list to identify the, subjects in the particular experiment being planned and will choose from other lists to identify the type of cells in the set of samples to be tested for each subject. The user selections will then be transferred to the sample treatment protocol, where the reagents for each sample and the treatment protocol will once again be specified by selecting from lists of treatments, etc.

Once the information for each protocol is complete and the user specifies readiness to begin work on the experiment, the collected information will be processed to create an XML file, which will then be transformed by an XSLT style sheet and passed (via a local proxy server)to the central LDAP server. Later, this process will be "reversed" and a copy of the information relevant to FACS data collection will be passed to the data collection modules. After data collection, collection-related information (including the location of the data that was collected), will be processed into an XML file, transformed by XSLT and sent to the LDAP server to be linked with the original protocol information, which will then be used to support data retrieval and analysis.

The programming for the GUIs and the XML >XSLT >LDAP linkages present no particular problems and should readily be accomplished. The development of the fields for each GUI, and the lists of items supplied for each field, however, will only be minimal when the first test users begin using the interfaces, since only a subgroup of the potential users will have had the opportunity to add items to the various selection lists. These lists

will be enlarged as more sites are added and will likely continue to grow when the system is released for broad usage. At some point, if the lists become too cumbersome, we will consider methods for making sublists available to users.

We will create builders for protocols that capture three types of information from investigators

a) **Study protocols** capture the hypotheses to be tested and the factors that go into them, including subjects, treatments, experiments and the timeline for an overall study. For a simple study, this means the subjects, the treatment and the end-point measurements. However, for a more realistic and complex study, there will be many potential factors and measurements and likely several hypotheses. In fact, particularly in basic science studies, investigators will probably find it necessary to refine and extend this definition during the "discovery" process.

The value of capturing this information early, and at the highest level of abstraction, is that IBRSS Wizards will then be able to automate much of the tedious "cut and paste" of data between various programs, which is extremely error prone and consumes an immense amount of the investigators time. In addition, and perhaps most important, this information will drive the "statistical consultant" Wizard and enable it to configure a statistical platform suitably for the data presented, thus relieving the investigator of having to answer "statistics" questions about data format, type and other issues that usually confuse the average biomedical scientist.

b) **Experiment protocols** capture annotation information to define the subset of subjects for which data will be collected, the set of samples to be obtained from the subjects, and the analytic procedures and data collection instruments used to analyze the samples.

c) **Sample-treatment protocols** capture annotation information to define the subdivision (aliquotting)and the treatment (reagents and conditions)for a set of samples for which data will be collected by a single analytic method (usually a single instrument).

2. Develop novel methods for automatic aspects of protocol specification (build Wizards)

Different parts of the protocol information will be used at different points in the

data flow, e. g., sample preparation or staining, FACS data collection, FACS data analysis, and experiment or meta analysis. However, while users have a clear concept of these processes, they rarely have a sense of the data flow underlying the experiments they perform. Therefore, even if they were willing to take the time to play "twenty questions" with a protocol builder, they would be unlikely to be able to provide the information necessary to take full advantage of automation to facilitate data collection and analysis. In particular, they would be hard pressed to understand the statistics jargon in which the questions are couched (biologists have enough jargon of their own to hande).

The trick here is to structure the user interface such that the easiest way for the user to enter information about the experiment will provide the cues needed concerning the structure of the data. For example, since it is easier to enter two variables that are to be crossed than to fill out a whole table by hand, users can readily be convinced to simply enter the two variables and leave the crossin (filling out the final protocol table)to the technical or statistics assistant. This point is illustrated in the example that follows:

To find ways to encourage users to take full advantage of the protocol builder's capabilities, we will devote a portion of our effort to more free-ranging research in which we will explore some novel methods that we believe can automate tedious protocol-specification tasks that are not generally considered to be amenable to automation. Initially, we will focus on the following:

a) **Capturing the model**

In a simple mouse experiment, for example, the subjects will typically be a number of individuals from an inbred strain, i. e., nominally identical, so they don't need to be randomized. The user may want to immunize with protein X, protein Y or nothing and then treat or not treat the animals in some way, e. g.,with UV irradiation. The user defines three attributes: mouse strain which has values Strain B and Strain C; immunization, which has values X and Y or nothing; and treatment, which has values treated or untreated.

Statistically speaking, these are three crossed nominal variables, but we probably should not try to convince the user of that. Instead, we should capture the model from the information the user enters, i. e., we should set up the interface such the user can choose to enter the three attributes and their possible. values and cross them rather than fill out a 12 row by 3 column (36 cell)table in which all combinations are accounted for. This would clearly be less work and certainly less error prone.

b) **Automating definition of plate/rack layouts and creation of reagent "cocktails"**

A technical assistant (Wizard)that provides worksheets for dilutions and cell counts and could identify and/or schedule the appropriate (FACS)instrument to use for data collection. It might use combinatorial methods to identify feasible combinations of available reagents and might then rank them by cost or power (by a process yet to be defined). It could also provide layout assistance and might customize the user interface to deal with different classes of experiments (e. g., mouse vs. human).

In experiments where more than one reagent is added per tube (or well), investigators may find it easier to pre-mix the reagents and do only a single addition of a reagent "cocktail". When only a few reagents are involved, this is a rather simple process requiring only that the Wizard calculate the total amount of each reagent to be added to the cocktail and the amount of the cocktail added to the tube to maintain the appropriate final reagent concentrations during the incubation. However, when the number of reagents to be added is large (e. g., 11-color FACS work requires cocktails that may include over 20 reagents, several labeled with the same fluorescence "color" and most labeled with distinct fluorescence colors), a Wizard can greatly facilitate the construction of the cocktail by providing a worksheet that keeps track both of the reagent and its color and assures that the desired combinations are reached.

c) **Checking for inadvertent omission of controls and automatically supplying controls**

Inexperienced (and even experienced)investigators quite frequently find they have to repeat experiments because they have not included "negative" controls that report the autofluorescence of unstained cells or the amount of second-step ELISA reagent that binds in the absence of first-step antigen-specific reagents. By developing Wizards that can be "told" what kinds of controls a particular laboratory wants to include in experiments of a given type, the Wizard can readily "suggest" addition of controls and add them if the "suggestion" is accepted. Complex control set-ups may require special Wizards. However, for the basic types of controls in most experiments, a very simple set of Wizard capabilities should suffice and should result in significant savings of reagents, samples and investigator time.

d) **Facilitating data collection and analysis**

Instrument control. The ScienceXchange model foresees the use of information captured by protocol builders to facilitate data collection, to permanently associate protocol information with data as it is collected, and to pass necessary information to analysis packages for statistical procedures and for labeling axes in graphs and column heads in tables. In addition, appropriate information can be displayed for each sample during manual data collection. For automated data collection, information entered at the protocol stage can drive the data collection, including specification of analysis parameters (how much, how many, how long)for individual samples and for the whole analysis. We are not requesting funding here for construction of data collection or analysis modules that could provide these capabilities for various instruments, since such modules usually must be built in collaboration with instrument manufacturers. However, we plan to design all protocol builders and their Wizards with the ability to collect and export the necessary information and have already begun recruitment of instrument manufacturers to develop appropriate modules (see section 3, below).

Analysis and data visualization. To provide a concrete example of the ways in which an analysis Wizard may work, we once again return to the rich and complex data source (and source of user difficulties)that FACS provides. Analyzing FACS or similar data and seeing the results of the experiment involves two processes, one of which is literally analytic in the sense of dividing up (gating to define subsets)and the other is essentially visualization, graphics and enumeration ("seeing" the data). The user switches back and forth between them and thinks of the whole process as "FACS (or other) analysis".

A Wizard could use information entered at the protocol stage to automatically call for processing (analysis) of FACS data from simple experiments like controls or titrations. Sometimes, this processing would occur as soon as the data is collected; at other times, gating or other information would first have to be obtained from the user. The Wizard might also suggest appropriate ways to visualize specified sub models based on the number, type and cardinality of the various factors.

Visualization of the *FACS data sets* (cell data)associated with *FACS sample data sets* and sub *sets* is the second major component of FACS analysis. Visualization tools are used to view FACS data, to define the polygons used to compute the Boolean (gating)functions, to reduce FACS data to interpretable results and to produce publication graphics. Typically, axis labels and legend information on the visualized graphics are

constructed by associating the color of each FACS measurement (raw or compensated)with the value of the attribute of the same name inherited from the stained sample, e. g., CDllb labeled with flourescein. Scale information is taken come from instrumentation values recorded by the collector; other values may come from other components of the system. Visualization may be used during data collection, analysis or even during construction of the protocol.

Programs such as FlowJo (TreeStar, Inc., San Mateo, California)and CellQuest (Becton-Dickinson Biosystems, Milpitas, California)provide both of these capabilities. FlowJo currently accepts axis labeling and other information from FAGS/Desk and will accept similar input from ScienceXchange. Hopefully, future FlowJo versions will also accept output from the ScienceXchange Wizards (discussion are underway toward this end; see section 4, below).

To complete the circle, ScienceXchange will accept output from analysis and other programs and enable storage of this output in an organized fashion, together with the raw data and other relevant information. We will make provision for "drag and drop" input of analysis information (and of export of raw data for analysis)to and from third party analysis software vendors, but again, we hope to develop closer cooperation with such vendors. Basically, we would prefer that definitions of the graphics be storable in fairly abstract form so that they can be rendered (or rerendered)locally according to the capabilities of the user computer and the user's preference. This would allow an analysis assistant, for example, to automatically generate a report containing a graph of a given type for each class of particular classifier attribute, or for a pre-specified set of classes of the classifier.

Statistical treatment and overall summaries. Having found and isolated several subsets, and obtained frequency and other information about these subsets, users are likely to have to summarize this information into something comprehensible. Typically, for some or all populations, the user will define a new computed attribute for some samples that corresponds to the frequency (or absolute count) of some population, or the mean or percentile of some measurement over a population. This may require normalization with some other data source. For example, in the clinical trial, FACS data provides the frequency of CD4 T cells, but the desired output value is the (absolute)number of CD4 T cells per microliter of blood. This is computed as absolute CD4 =CD4 Lymph /Total Lymph *Lymph Concentration, where the first two factors are FACS frequencies and the last is a clinical lab result.

Although Wizards can readily be organized to do these types of computations, and although the study model can readily specify the needed work, there is little help at present for automating these crucial data summary operations. For example, FlowJo outputs tables of computed data (mean fluorescence, frequency, etc.)for various subsets that the user identified. At present, these tables can be imported into Excel for further processing. Alternatively, many users import them into JMP, a statistical discovery program developed and marketed by the SAS Institute (Cary, North Carolina; see section 4 below). However, in either case, it would be much more efficient for the user if ScienceXchange Wizards were to manage the data export to FlowJo (or other computation packages) and were to accept FlowJo output, which could then be sent to JMP (or other statistics packages) along with the clinical lab values or other values necessary to obtain the final analysis results (which, after all, are what the user is looking for).

This scenario would also allow the user to capitalize on additional information at the study level and would enable appropriate testing of the hypotheses defined for the study. Thus, armed with output data for each *FACS data set* or *subset,* the user is in a position return to the summary *FACS experiment data to test* hypothesis concerning the impact of the experiment variables. For the mouse experiment, assume that the user has defined two populations (say T-cells and B-cells)and measured the median fluorescence for CD45 (or another cell surface antigen)on each of them for all samples. If the user selects mouse strain, *immunization, treatment* and median CD45 of T-cells, the statistics expert deduces that this sub model has independent crossed nominal variables *(mouse strain, immunization* and *treatment)*and a continuous dependant variable (the median CD999 fluorescence). This allows the expert to configure the IMP ANOVA platform to test the hypothesis that the treatment or the priming or both had some effect on that population (increasing or decreasing median CD999 fluorescence). Selecting both medians would configure a MANOVA platform. Selecting an independent time variable might launch a time series specific platform, etc. These platforms are available in callable statistics packages such as JMP, which also generate graphical output. Selecting out the data from BALB mouse strain, rather than all mice in the experiment, produces a sub model with *immunization* crossed with *treatment* as the sub model and the same dependant variables.

End Confidential

At present, users are required to interact with at least three distinct software

packages and to have a strong grasp of the pitfalls of data collection and statistical analysis to successfully navigate the above. The ScienceXchange mission, to be achieved in part by work proposed here and in part by work supported in other ways, is to produce an overall system that will support accomplishment of research goals with considerably less struggle.

3. **Co-operate with instrumentation manufacturers to develop data collection modules that utilize the protocol information captured by ScienceXchange (no support requested)**

As indicated above, it is useful to pass protocol information to the data collection module to inform data collection and assure that the collected data is properly associated with the protocol and study information to facilitate analysis. This process can be made to operate without cooperation from instrument manufacturers provided that users intervene to associate the data file collected for a given sample with the protocol information for that sample. However, we intend to seek cooperation with instrument manufacturers to integrate data collection more closely with ScienceXchange capabilities.

The development of integrated data collection with FACS instruments will serve as the model for integrating data collection with other instruments. Until the protocol builders are in place, Stanford and the two beta test sites continue to depend on FAGS/Desk, both for entry of protocol information and for actual data collection. Since maintenance of this archaic program is tenuous, there is strong motivation to rapidly develop relatively simple interfaces that will replace the functionality of the older modules and thus (finally)enable complete abandonment of FAGS/Desk. We have already secured verbal agreement (letter to follow)from Becton-Dickinson Biosystems (Milpitas, California), the manufacturer of the FACS instruments at these sites, to co-operatively develop a data collection module that will use information gathered by the protocol builder to inform data collection and will send the collected data, appropriately associated with the protocol information, to ScienceXchange. Similarly, we have secured verbal agreement from Gene Machines, Inc., to adapt a protocol interface to their DNA microarray spotter.

We are not requesting support here for this aspect of the project.

4. **Create tools to enable launch and use of third-party analysis and visualization programs**

In essence, to allow users to take full advantage of the potential inherent in the

protocol information capture mechanisms, ScienceXchange will have to either create analysis and visualization packages capable of utilizing this information or arrange cooperative development with third party software vendors who want to capitalize on the market that these capabilities address. Our experience to date suggests that vendors will readily be found for this purpose. As indicated above, a path has already been developed that enables passage of protocol information (axis labels, etc.)to FlowJo and discussions are in progress to enable passage of additional information and acquisition of FlowJo output into ScienceXchange. In addition, our discussions with Becton-Dickinson concerning data acquisition will also extend to developing an interactive route for work with their analysis package (CellQuest).

Finally, we have begun discussions with John Sall, Senior Vice President and Founder of SAS Institute and the lead developer of the SAS/JMP statistics analysis and discovery software, concerning development of modules that will enable ScienceXchange to import and export of data in JMP tables. Hopefully, we can report the success of these and the above discussions before this proposal is reviewed.

5. **Establish test sites**

a) **Maintain the current alpha test site at Stanford University (Herzenberg laboratory)**

As indicated throughout this proposal, the IBRSS technology is based on prototype server software to be licensed from Stanford. The Herzenberg laboratory has hosted this development and will continue to host further server development, including establishment of the IBRSS alpha test version(s).

b) **Establish two beta test sites at start of Phase II**

As indicated above, the IBRSS alpha test site will be the Herzenberg laboratory (Genetics Department, Stanford University School of Medicine), where FAGS/Desk and the prototype for IBRSS was developed and where the initial IBRSS developer, Wayne Moore, is still employed as the senior software engineer. The first two beta sites will be located at Fox Chase Cancer Center (under Richard (Randy)Hardy's direction)and at the University of Iowa School of Medicine sites (under Morris Dailey's direction), These sites were chose because they have a currently operating FAGS/Desk installation.

Initially, we plan to import the FACS data archives at these sites into IBRSS and make the entries available to users over the Internet. As at Stanford, users will continue to

collect FACS data and protocol information with FAGS/Desk until their FACS instruments are retrofitted with interfaces that allow collection of FACS data directly into IBRSS and we complete at least a primitive replacement for the FAGS/Desk protocol utility.

All three test sites have full-spectrum research capabilities and are conducting internationally-recognized studies generating information of importance to many areas of clinical and basic importance, including lymphocyte development, bone marrow transplantation and a variety of issues relevant to the origin and control of neoplasia. FACS is a central tool for this research. However, the FACS work is embedded in studies that utilize a wide variety of instrumentation, ranging from imaging to DNA microarray spotters and scanners. Therefore, these sites provide the ideal setting for establishing and expanding IBRSS capabilities to provide integrated support for data intensive research of all kinds.

These sites are also useful because users tend to be accustomed to working with advanced prototype software. The Stanford site, in particular, has users who have pioneered the use of various FAGS/Desk capabilities and provided the alpha test site for FlowJo software, which was partially developed under Wayne Moore's supervision before migrating out into the commercial world. Investigators in the Herzenberg laboratory are thus trained to report bugs, find workarounds and generally co-exist with alpha level software. They are anxious to move to the IBRSS system despite this experience and look forward to commercial alpha support rather than the developer support that has been available to date.

Fox Chase and the University of Iowa adopted FAGS/Desk as a University to University exchange and have been accustomed to operating this system without any formal support. They also were among the first to adopt FlowJo and have therefore become accustomed to bug reports and workarounds. Like Stanford, investigators at these sites are anxious to move to IBRSS and are willing to put up with the inconvenience of beta testing.

This said, ScienceXchange looks forward to getting IBRSS working, first at the alpha and then at the two first beta sites, with minimal disruption of the work flow at the sites. The IBRSS prototype that we will import from Stanford has been in operation for some time and is largely debugged. We will add new capabilities and therefore expect some initial problems. However, we expect that these will mainly be ironed out by internal testing, making even the alpha test relatively trouble free.

c)&d) **Establish additional beta test sites later in Phase II**

We have already begun discussions with the director of a multi-center clinical cancer research consortium with the hope that ScienceXchange will soon be in a position to facilitate the collaborative research being carrier out at the seven sites in this consortium. Once IBRSS is operating smoothly and has linked in a number of different types of data collection instruments, we will recruit either this or another such consortium as an advanced beta site specifically designed to test IBRSS capabilities in enabling data sharing and other currently problematic aspects of joint studies.

We also plan to establish European and Japanese beta sites, to enable internationalization of IBRSS protocol builders and other services and to test the IBRSS capabilities on Internet sites outside the US. We have begun discussions toward this end with the director of a Spanish laboratory with multiple FACS instruments and are completing arrangements with a Japanese Instrumentation company (Tomy Instruments, Tokyo)to represent ScienceXchange in Japan and to locate two beta test sites there.

Start *Confidential:*

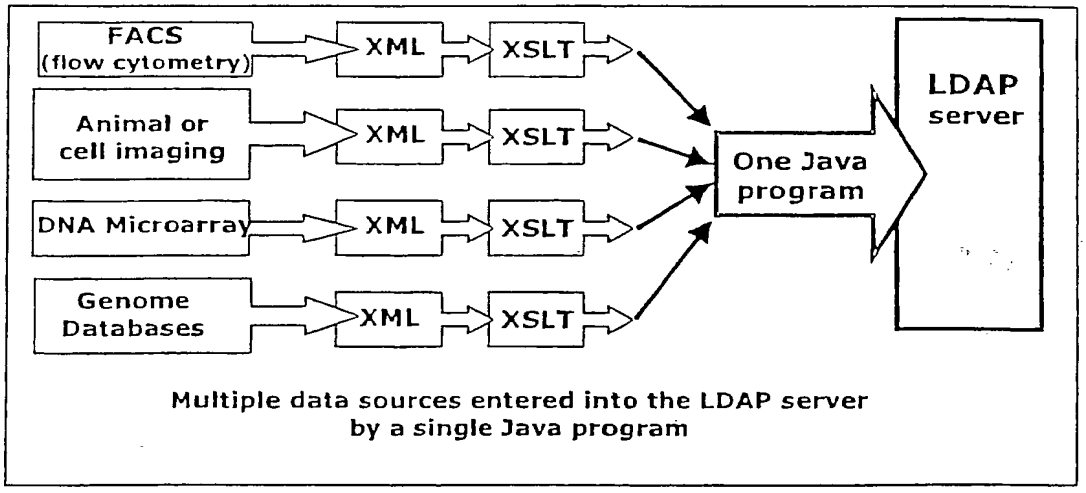
The following outlines the basic technology we plan to use to construct the IBRSS system. Although ScienceXchange is dedicated to the use open standards in IBRSS wherever possible, we have labeled this confidential because it includes specific notes defining the way we intend to use this technology.

Coda: LDAP directories in the service of biomedical studies

XSLT style sheets were developed to provide the information for rendering XML documents for viewing in browsers. However, recognizing that this transformation process is not restricted to rendering documents for viewing, ScienceXchange is putting it to unique uses in the scientific arena. In essence, the stylesheet transformation language (XSLT)defines the transformation of the original input (XML)document to "formatting objects" such as those included in HTML documents. In a traditional style sheet, these are then rendered for viewing. However, the XSLT transformation grammar can also be used to transform XML documents from one form to another, as in the following examples:

a) **Loading directories.** XSLT can be used to transform an XML file generated by any data processing application to an XML representation of a directory (sub)tree, i. e., to extracting directories entries from the XML document. The

ability to use XSLT for this transformation greatly simplifies the creation and maintenance of LDAP or other directories that serve diverse information derived from distinct sources (e. g, FACS instruments and genome data banks)that generate different types of XML documents. In essence, using XSLT removes the necessity for writing distinct Java code to construct the directory entries for each type of document. Instead, appropriate "directory styles" can be defined for each document type and a single Java program can be written to process all XSL-transformed documents into the directory tree (see figure).



Multiple data sources entered into the LDAP server by a single Java program

b) **Re-indexing directory entries.** Existing documents may be readily re-indexed based on any desired elements or attributes present in the XML documents simply by changing the XSLT style sheet. Changes in the directory schema may be required for extensive indexing changes but could also be driven by an XML representation of the appropriate schema.

c) **Cataloging new documents.** A new type of document can be cataloged simply by creating an appropriate XSLT style sheet and modifying the directory schema if necessary, as above.

d) **Cataloging from arbitrary XML documents.** A default XSLT directory style sheet can be created to extract a pre-defined set of indexing elements

included in arbitrary XML documents. This would enable creation of the corresponding directory entries for these indexing elements.

e) **Passing information from XML files to analytic or other programs:** XSLT can be used to transform a subset of the information in an XML file so that it can be read by a program that takes XML input in a particular format. In addition, XSLT can launch the program and pass the result of the transformation during the launch. For example, using XSLT stylesheets, we can launch an analysis application by transforming an XML file containing the results of a directory search to an application-readable file containing URLs for the data and appropriate annotation information for the analysis. This option can be made available for all co-operating applications and need not be restricted to FACS data.

f) **Creating data displays.** XSLT style sheets can be used to change the form of a document. For example, they can be used to extract the results of analyses and display them as values in the rows or columns of a table.

Storing analysis output: As indicated above, we plan to use XSLT and other capabilities to develop mechanisms for storing analysis output along with the primary data and annotation information. We have already begun conversations with third-party vendors of FACS analysis software about modifications to their systems to enable storage of their computed output. Alternatively (or in addition), we are will develop fully cooperating applications for analysis of FACS and other data. This function is central to the sharing of data interpretations and must be addressed to complete the deliverables of the project. Since our consultants in the Herzenberg lab are highly experienced in the development of analysis modules for FACS data, we do not view this as a particularly difficult problem.

Expanding directory searches. Together with Moore and the Herzenberg lab, we plan to explore additional search mechanisms that would allow "reversal" of the catalog process. At present, information is promoted upward from the documents into the directory for searching and no searching is done within the documents. However, since XQL allows searches to proceed downwards from the directory, ScienceXchange will investigate the possibilities of a search application that uses the LDAP search functions to retrieve a set of candidate XML documents (based on their directory attributes)and then uses XQL to further refine this set. To facilitate

XQL use, ScienceXchange will provide a unified interface that would largely make the differences in search strategies transparent to the user. If this approach proves feasible, the user will be able to select (search and retrieve)for items within the document that are not reflected in the directory or could extract elements from these documents, e. g., samples from a set of experiments.

User and instrument interfaces for collecting FACS data. Collection, transmission and storage of annotated data is central to the ScienceXchange mission. However, we do not have the laboratory and engineering facilities necessary for this development. Therefore, for FACS data, we will co-operate the Herzenberg lab and will establish similar co-operations for other instrumentation. When the applications are complete, we will license them for distribution to other users. We are not requesting funding here for this cooperative development effort,

Herzenberg lab has already developed plans to develop the structured mechanisms for collecting primary FACS data, for annotating it with information generated during the data collection, and for transmitting the annotated primary data to the ScienceXchange LDAP server for storage in association with the appropriate XML-encoded experiment and study descriptions. The following modules are planned:

a) **Set-up module(s)**-automate aspects of instrument set-up and standardization; record and visualize relevant instrument information; acquire and respond to user input

b) **Data collection module(s)**-collect primary (instrument-generated)data for the aliquots of each sample; visualize protocol information to facilitate data collection; acquire and respond to user input; record machine condition and user comments specific to each data collection.

   i) Where possible/permitted, adapt and interface the data collection modules to specific machines (e. g., various FACS, imaging and DNA-array data readers)to provide full functionality for data collection.

   ii) For instruments that do not provide/permit direct access to machine control and data collection, develop additional modules that enable manual entry of machine information and "point-andclick" association of primary data collected for each sample aliquot with the protocol information for that aliquot.

c) **Extension of the FACS document type** -include new functionality such as instrument setup, auto-calibrator and quality control elements, tabulated

transfer functions and operator commentary in the definitions of the FACS document type. Provisions for digests of the data files that are referenced and for digital signatures will also be made.

d) **Data transmission module(s)**-link (annotate)the primary data with protocol instrumentderived information; communicate authenticated (digitally-signed)primary data and its annotation linkages to the information store.

Storage of data and annotation information -ScienceXchange will develop a reliable, large scale (terabyte level), web accessible, central storage system coupled with small-scale volatile storage deployed locally in a manner transparent to the user. This system will store data and annotation information transmitted from the data collection system. In addition, it will catalog the stored data according to selected elements of the structured annotation information and will retain all catalog and annotation information in a searchable format. Wherever possible, ScienceXchange will use industry standard formats for storing data and annotation information. If no standard is available, ScienceXchange will publish the interim formats that are used and provide translators to industry standards that become available.

Federated directory and data storage -ScienceXchange will capitalize on the built-in replication and referral mechanisms that allow search and retrieval from federated LDAP networks in which information can be automatically replicated, distributed, updated and maintained at strategic locations throughout the Internet. Similarly, because pointers to raw data in LDAP are URLs to data store(s), we can capitalize on the flexibility of this pointer system to enable both local and central data storage.

Maintenance of data and annotation information security -ScienceXchange will enable highly flexible, owner-specified "fine-grained" access controls that prevent unauthorized access to sensitive information, facilitate sharing of data among research groups without permitting access to sensitive information, and permit easy global access to non-sensitive data and analysis results.

a) **Built-in access controls** that prevent release of unauthorized information from the system

b) **Multi-level access controls** to allow data owners to specify which users, or classes of users, are permitted to retrieve individual data sets and/or to access individual elements of the annotation information during searches

c) **User identity verification system** that is referenced by the access control system

d) **Anonymous access to data and annotation information** that owners make available for this purpose

Note: LDAP provides fine-grained security controls that give the individual user control over individual elements that will be exposed or hidden. However, the overall issue of security needs to be considered -from an Internet perspective. For example, we are currently grappling with the following: should the data be encrypted or the server or only on the wire? do we need to require (or allow)secure sockets for most operations? what sort of digital signatures, message digest and cryptography algorithms should we use?

Retrieval of data for analysis -ScienceXchange will enable retrieval of annotated data sets and transfer to visualization and analysis programs that can use the annotation information to label analysis output, facilitate data interpretation and enable return, storage and retrieval of analysis output within the context of the study and experiment that generated the primary data.

a) **Retrieve annotated data sets** (subject to owner-defined accessibility)via catalog browsing and/or structured searches of the catalog; automatically verify authenticity of the data based on the digital signature.

i) Launch internal and co-operating data analysis and visualization programs and transfer the data and annotation information to the program

ii) Put the data and annotation information into published-format files that can be imported into data analysis and visualization programs that do not provide launchable interfaces

b) **Retrieval analysis output** -recover/import and link analysis output with primary and annotation data to provide access to findings via subject and treatment information that was entered at the study and experiment levels.

i) **Store and catalog output** from co-operating analysis programs (within the limitations imposed by the capabilities of analysis programs that were not designed for this purpose).

ii) **Develop internal analytic modules/programs** that will enable users to fully capitalize on the annotation information entered into the system.

end Confidential

Additional uses of the information and data collected in IBRSS

Using the technology discussed above, ScienceXchange will build a federated web-accessible system that enables creation, cataloging and functional availability of standardized data sets. that can be utilized as a national repository of flow cytometric information. Once the beta testing is complete, IBRSS will be open to the scientific "public". This will bring a larger and more diverse group of investigators (and their data) into the system and thus will help to broaden the initial base from which addition changes to the protocol generators can be made. In addition, it will provide a wide variety of biomedical studies that could be made available as part of an overall program to make (owner-released)scientific data resources available over the Internet. Facilities in this program could provide the following:

a) **Repository for primary data abstracted in publications** -a resource to enable direct access to the primary data upon which display items (tables, graphs)in publications are based. The Federal government is considering mandating such access to primary data.

b) **Library of cell surface expression patterns for types and stages of disease** -a resource to enable researchers and clinicians to facilitate diagnoses and definitions of new conditions by comparing with locally acquired FACS and other data with resource data acquired from characterized subjects.

c) **Data source for science education projects** -a resource to provide science educators at all levels with standardized data that can be used to teach analysis, data interpretation and diagnosis methods. In addition, it will provide material for student research projects and for examinations.

Time Line for Phase II

During year 1, we will create the basic interfaces necessary to capture study, experiment and sample treatment annotation information. We will use FACS studies as the primary model for creating these protocol builders but will create them with a broad approach that will allow migration to other biomedical technologies.

In addition, we will complete the basic IBRSS server technology and open it to researchers at the Stanford alpha test site.

By the end of this year, we expect to be routinely moving annotation information

and FACS data from the Stanford site to the ScienceXchange IBRSS site and serving that information and data to researchers at the Stanford site. We will also develop launch capabilities for at least one FACS data analysis and visualization program (most likely FlowJo).

During year II, we will develop novel methods for automatic aspects of protocol specification that will facilitate data collection and analysis by enabling capture of the experiment model, automation of the definition of plate/rack layouts and creation of reagent "cocktails", checks for inadvertent omission of controls, and automated suggestion of omitted controls. We will also open two beta test sites (Fox Chase Cancer Center and the University of Iowa)during this year.

Finally, and perhaps most important, we will begin expanding IBRSS to acquire and serve annotation information and data from other biomedical technologies and will make these capabilities available to the alpha and beta test sites.

During year III, we will continue with the diversification of IBRSS capabilities. In addition, we will open several addition beta test sites to enable IBRSS internationalization and to expand IBRSS to support cooperative basic and clinical research at multiple centers.

We will clearly be pleased if we can speed up this timeline and begin reaching year II and year III goals prior to the formal start of these years. If so, we will have more time in year III to expand IBRSS capabilities and adapt it to supporting co-operative clinical research in cancer and related fields.

Testing and Evaluation

The completion of Phase II will require completion of working prototypes of the three interfaces listed in Specific Aims. To meet the criteria for a working prototype, each protocol generator will have to be able to 1) present lists of standardized choices that collectively enabled intake of the annotation information necessary for the study; 2) record user selections; 3) provide "type-in" capabilities for items that are not amenable to listing; and, 4) provide the ability to transfer the acquired annotation information to the archive index or to a central "information distributor" in the overall system, .e. g, for transfer of the relevant components to a data collection module that provides access to certain annotation information during data collection.

The code produced to meet Phase II goals need not be fully optimized but must be stable enough for beta testing and thus must allow repeated use without crashing. Further, mechanisms for selecting reagents and other types of standardized annotation produced in

Phase II must be fully operative but need not provide a complete range of options. Year I will be devoted to determining as many of these options as are deemed useful by the restricted group of alpha testers (Herzenberg laboratory scientists)with whom we will work during this Phase. This list will be extended during year II as the beta test process brings us into contact with a substantially broader group of investigators.

The criteria for completion of the evaluation of the novel methods listed above (Goal 2a-d)is somewhat different. The listed goal will be considered achieved (complete)either if a method is designed and successfully incorporated into a working prototype or if test data or feasibility studies rule out further exploration of the method.

Years II and III of this project will be devoted to installing the protocol generators into the overall ScienceXchange system and beta testing them together with the overall system at several academic and research institute sites. Completion will require successful correction of errors recognized during the beta test and extension of the protocol generators to serve the needs of the broad cross-section of biomedical researchers working at the various beta test sites.

begin confidential

LDAP object classes

The tables that follow provide some examples of how objects will be represented in the IBRSS LDAP directory. We have marked these tables confidential. However, when IBRSS is implemented, they will be made public along with all other standards in the directory.

Table I: Scientific Investigator

objectClass	cis	ScientificInvestigator, InetOrgPerson, organizationalPerson, person
UID	cis	User identifier must be unique in context
ou	cis	From distinguished name
o	cis	From distinguished name
professionalName	cis	Author name(s) used in the literature
professionalSpeciality	cis	For example "Cellular Immunology"
professionalAffiliation	cis	For example, "National Academy of Sciences"
professionalPublication	dn	ScientificPublication of which this is an author.

Table 2: Scientific Instrument

objectClass	cis	ScientificInstument
cn	cis	Common name
ou	cis	From distinguished name
o	cis	From distinguished name
instrumentManufacturer	dn	For example, ou=immunocytometry Systems, o=Bectorn Dickenson
instrumentModel	cis	For example, "FACS-II"
instrumentSerialNumber	cis	Manufactures id
responsiblePerson	dn	Dn of a person responsible for the instrument

Table 3: Scientific Publication

objectClass	cis	ScientificPublication, document
title	cis	Title
volume	cis	Volume
ou	cis	From distinguished name
o	cis	From distinguished name
pages	cis	Range of pages
reference	dn	Distinguished name of publication referenced by this publication
citation	dn	Distinguished name of a publication which referenced this one
author	dn	Distinguished name of author

Table 4: Monoclonal antibodies

objectClass	cis	MonoclonalAntibody
clone	cis	Unique clone name
o	cis	From distinguished name
ou	cis	May be part of distinguished name
UID	cis	May be part of distinguished name
cn	cis	Common name(s)
specificity	dn	Distinguished name of specificity
creatorDn	dn	Distinguished name of person or organization that created the clone.
titre	float	
concentration	float	
manufacturer	dn	Designated name of manufacturer
heavyChain	dn	dn of heavy chain locus or allele
lightChain	dn	dn of light chain locus or allele

Table 5: FACS instrument

objectClass	cis	FlowCytometer, scientificInstrument
cn	cis	Common name
instrumentManufacturer	dn	For example "ou=Immunocytometry Systems, o=Becton Dickenson"
intrumentModel	cis	For example, "FACS-II"
instrumentSerialNumber	cis	Manufactures identifier

Table 6: FACS experiments

protocolIdentifier	cis	Uniquely identifies protocol in context
UID	cis	May be part of distinguished name
instrument	cis	May be part of distinguished name
o	cis	May be part of distinguished name
ou	cis	May be part of distinguished name
instrumentDn	dn	Distinguished name of a scientific instrument
archiveURL	url	URLs of archive file corresponding to this experiment
dateCollected	date	
numberOfSamples	int	Number of samples collected

Table 7: FACS sample

protocolCoordinate	cis	Uniquely identifies sample in protocol
protocolIdentifier	cis	Uniquely identifies protocol in context
UID	cis	May be par of distinguished name
instrument	cis	May be part of distinguished name
o	cis	May be part of distinguished name
ou	cis	May be part of distinguished name
cn	cis	Common name
title	cis	Experiment title
description	cis	Description of the sample
sampleLabel	cis	Label for the sample from the protocol
investigatorDn	dn	Distinguished name of the investigator responsible for collecting the data
instrumentDn	dn	Distinguished name of a scientific instrument
dateCollected	date	
startTime	time	
endTime	time	
numberOfMeasurements	int	Number of components measured for each event
numberOfEvents	int	Number of events in the sample
URL	url	URLs of date file corresponding to this sample

***end confidential*

WHAT IS CLAIMED IS:

1. A method for managing a database using a lightweight directory access protocol for identifying and storing information with said database, said method comprising:

(a) applying said light directory access protocol to create a directory structure for said database, said directory structure having a plurality of nodes with distinguished names;

(b) defining standardized data; and

(c) transforming said standardized data for mapping onto said plurality of nodes.

2. The method of claim 1, wherein said step of defining comprises annotating said standardized data.

3. The method of claim 1, wherein said transformation step further comprises:

(a) developing extensions with XML; and

(b) mapping said extensions to said plurality of nodes.

4. The method of claim 3, wherein said developing step further comprises adding to said extension elements selected for the group consisting of cross-references, external pointers and links.

5. The method of claim 1, further comprising a centralized Internet-accessible archive for storing, analyzing, retrieving, and sharing said data.

6. The method of claim 1, further comprising a security means for user-controlled sharing of the data.

7. The method of claim 1, further comprising a structured hierarchy, said structured hierarchy comprising, in order:

(a) studies;

(b) experiments;

(c) data; and

(d) analysis.

8. A method for managing supplies of a laboratory using a computer comprising:

(a) Recording supplies used in said laboratory on a computer readable medium;

(b) Recording amount of said supplies in said laboratory on a computer on said computer readable medium; and

(c) Notifying laboratory personnel of said amount of said supplies.

9. The method of claim 8, further comprising sending automatic notifications when said amount of supplies is low.

10. The method of claim 9, wherein said automatic notification is performed through electronic mail.

11. The method of claim 8, wherein amount of said supplies is automatically updated.

12. The method of claim 11, wherein:

(a) said computer is coupled to an experiment protocol tracker;

(b) said automatic updating of said supplies occurs when an experimental protocol on said experimental protocol tracker is completed.

13. The method of claim 12 wherein, said automatic updating further comprises an estimation of supplies wasted.

14. The method of claim 13, wherein said estimation of wasted supplies comprises estimating supplies wasted during an experimental protocol.

15. The method of claim 8 wherein said computer readable medium comprises a database using a directory access protocol for identifying and storing information with said database.

16. The method of claim 15,wherein said database comprises:

(a) applying said directory access protocol to create a directory structure for said

database, said directory structure having a plurality of nodes with distinguished names;

(b) defining standardized data; and

(c) transforming said standardized data for mapping onto said plurality of nodes.

17. A method for interactively creating experimental protocols comprising:

(a) a user selecting options from a plurality of experimental options;

(b) said experimental options being stored in a computer;

(c) said computer provides additional options to the user based on said user selected options.

18. The method of claim 17, wherein said interactive protocol maker is coupled to analysis software.

19. The method of claim 18, wherein said analysis software is further used to provide said experimental options to the user.

20. The method of claim 19, wherein proper units are automatically transferred to the analysis software.

21. The method of claim 17, wherein said additional options are based on the current inventory of laboratory supplies.

22. The method of claim 17, wherein an output comprises an experimental protocol comprising:

(a) instructions detailing the steps of the experiment to be performed;

(b) supplies to be used in the experiment; and

(c) instructions detailing which of said supplies and which subparts of said supplies are to be used during each of said steps in said experiment.

23. The method of claim 17 wherein said created experimental protocols are recorded on a computer readable medium comprising a database using a directory access protocol for identifying and storing information with said database.

24. The method of claim 23, wherein said database comprises:

(a) applying said directory access protocol to create a directory structure for said database, said directory structure having a plurality of nodes with distinguished names;

(b) defining standardized data; and

(c) transforming said standardized data for mapping onto said plurality of nodes.

25. A method for interactively creating study protocols comprising:

(a) inserting a hypothesis to be tested;

(b) inserting research criteria to be followed; and

(c) inserting research parameters to be followed.

26. The method of claim 25 wherein said created experimental protocols are recorded on a computer readable medium comprising a database using a directory access protocol for identifying and storing information with said database.

27. A method for interactively collecting data comprising:

(a) creating an interactive experimental protocol; and

(b) conducting experimental steps from said protocol.

28. The method of claim 27, further comprising storing data created from said experimental steps onto a computer readable medium.

29. The method of claim 28, wherein said computer readable medium is accessible over an interconnection network.

30. The method of claim 29, wherein said interconnection network is the Internet.

31. The method of claim 28 wherein the computer readable medium comprises a database using a directory access protocol for identifying and storing information with said database.

32. The method of claim 31,wherein said database comprises:

(a) applying said directory access protocol to create a directory structure for said database, said directory structure having a plurality of nodes with distinguished names;

(b) defining standardized data; and

(a) transforming said standardized data for mapping onto said plurality of nodes.

33. A method for analyzing data comprising:

(a) accessing experimental data from a computer readable medium;

(b) accessing an analysis program from a computer readable medium capable of analyzing said experimental data.

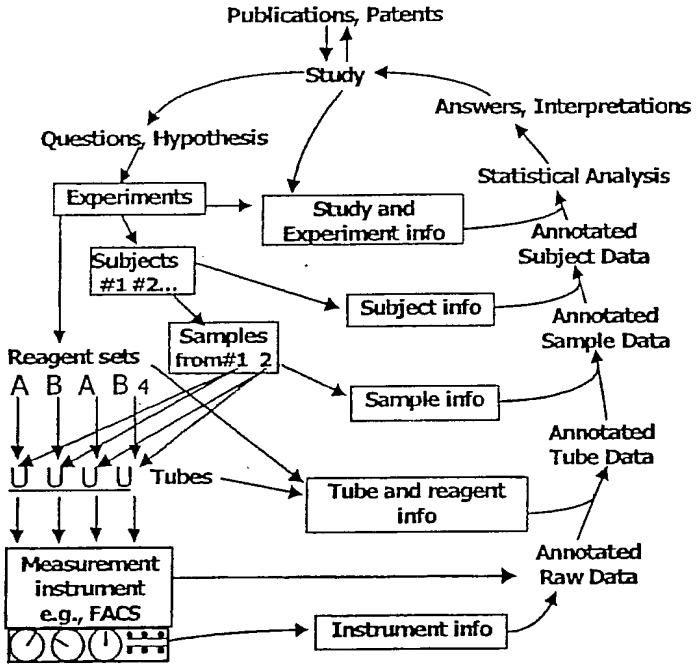
34. The method of claim 33, wherein the computer readable medium comprises a database using a directory access protocol for identifying and storing information with said database.

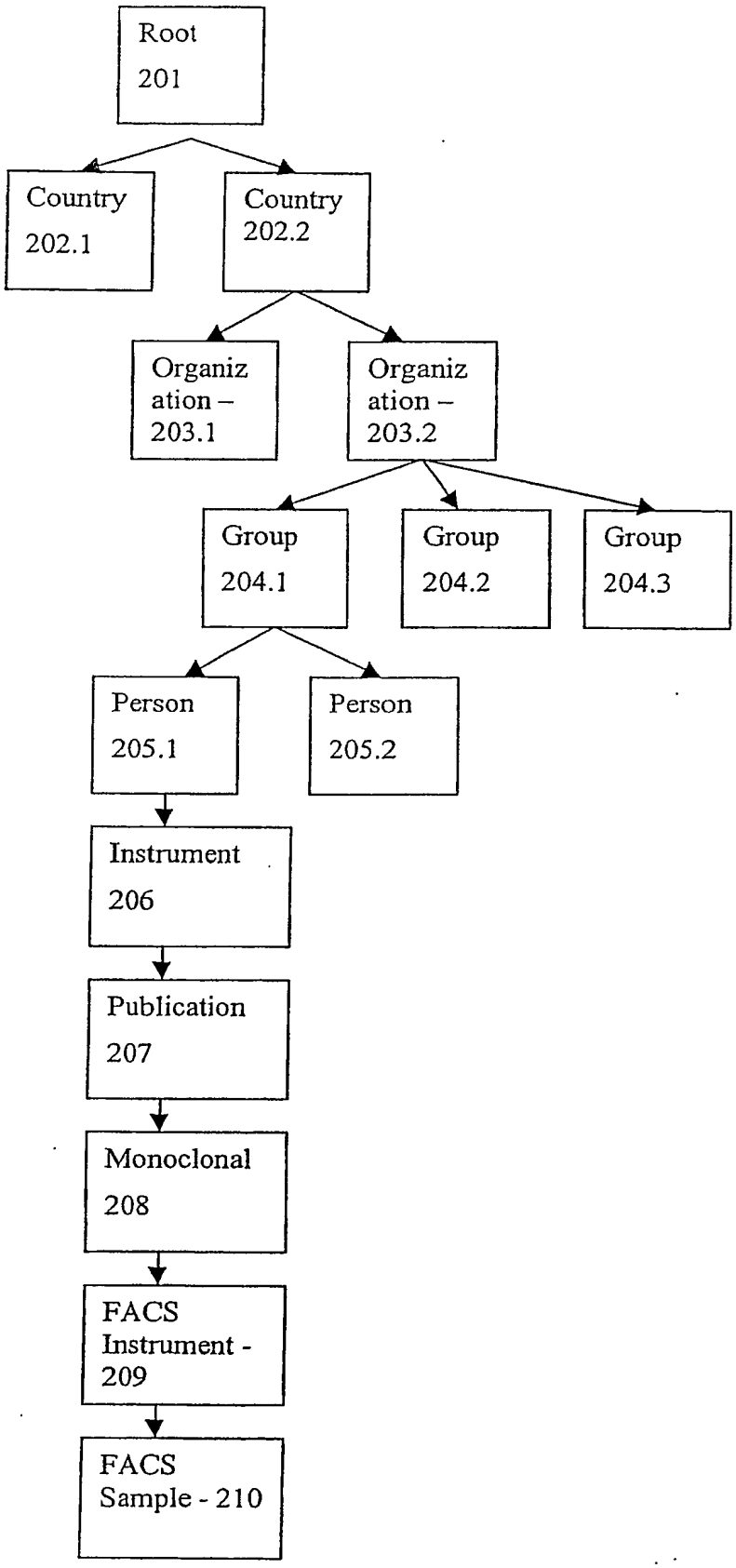
35. The method of claim 34, wherein said database comprises:

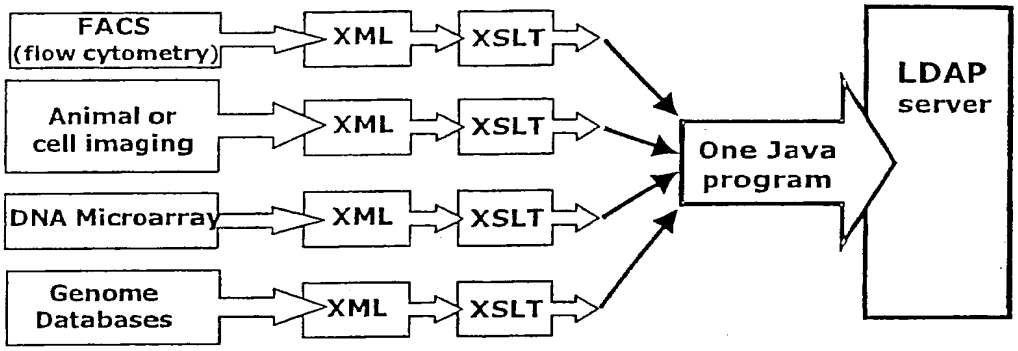
(a) applying said directory access protocol to create a directory structure for said database, said directory structure having a plurality of nodes with distinguished names;

(b) defining standardized data; and

(c) transforming said standardized data for mapping onto said plurality of nodes.


Publications, Patents
Study
Answers, Interpretations
Questions, Hypothesis
Statistical Analysis
Experiments
Study and Experiment info
Annotated Subject Data
Subjects #1 #2...
Subject info
Annotated Sample Data
Samples from#1 2
Reagent sets
A B A B 4
Sample info
Annotated Tube Data
Tubes
Tube and reagent info
Measurement instrument e.g., FACS
Annotated Raw Data
Instrument info




FACS
(flow cytometry)
Animal or
cell imaging
DNA Microarray
Genome
Databases
XML
XML
XML
XML
XSLT
XSLT
XSLT
XSLT
One Java
program
LDAP
server